PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12Q 1/68 | A2

(11) International Publication Number: WO 00/15847

(43) International Publication Date: 23 March 2000 (23.03.00)

(21) International Application Number: PCT/US99/20449

(22) International Filing Date: 8 September 1999 (08.09.99)

(30) Priority Data:
60/100,030 11 September 1998 (11.09.98) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 60/100,030 (CON)
Filed on 11 September 1998 (11.09.98)

(71) Applicant *(for all designated States except US)*: GENE LOGIC, INC. [US/US]; 708 Quince Orchard Road, Gaithersburg, MD 20878 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: STEWARD, Keith, Leroy [CA/US]; 13501 Giant Court, Germantown, MD 20874 (US). SHI, Qin [CN/US]; 10307 Royal Woods Court, Gaithersburg, MD 20879 (US). CARIASO, Michael, Contento [US/US]; 12652 Grey Eagle Court #12, Germantown, MD 20874 (US).

(74) Agents: FENSTER, Paul et al.; c/o Anthony Castorina, Suite 207, 2001 Jefferson Davis Highway, Arlington, VA 22202 (US).

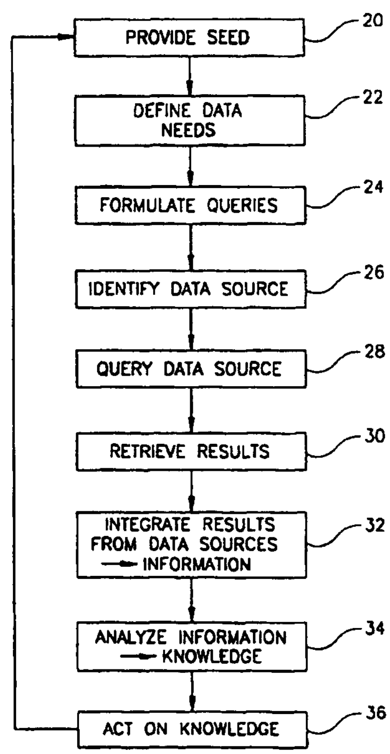
(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: GENOMIC KNOWLEDGE DISCOVERY

(57) Abstract

A method of genomic data discovery, comprising: (a) providing a gene data base comprising at least 10 genes; (b) selecting one of said at least 10 genes; (c) discovering knowledge for said selected gene; (d) repeating said (b) and (c) for a plurality of said genes; and (e) repeating said (b)–(d) a plurality of times such that knowledge is discovered substantially in parallel for all the selected genes. Preferably, (b)–(e) are performed substantially without human intervention. Preferably, knowledge discovery utilizes a large number of databases and inference rules to analyze data queried from the databases.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

GENOMIC KNOWLEDGE DISCOVERY

FIELD OF THE INVENTION

The present invention relates to automated knowledge discovery and, in particular, to knowledge discovery and data retrieval of genomic related information.

BACKGROUND OF THE INVENTION

Genomic research is currently an expanding field of endeavor. It is expected that by the year 2003, the entire human genome will be mapped out, listing an expected 100,000 different genes. One important customer of this data is the pharmaceutical industry, typically for use as "drug leads". When a new drug is to be developed for a disease, one approach is to discover a gene which interacts with the disease and, then, design a drug which modifies the disease by affecting the gene or proteins generated by the gene.

In earlier days of genomic research, genes were discovered by sequencing expressed mRNA fragments and using these fragments to identify gene locations, using biological methods. Once the gene was located, various methods, usually biological, were used to sequence the complete gene and determine its biological function. The relative amount of available biological data was quite small, so that cross-correlation between data sources was not performed often. These tasks took a considerable amount of time.

In recent years, many methods have been developed for (automatically) generating large amount of biological data, including for example, automatic analyzers, large scale gel-chromatography, high throughput expression profiling and DNA chips. In addition, the amount of research (e.g., micro-biology, neuro biology and structural biology) is also increasing. It is estimated that the volume of data currently doubles every 1.2 to 5 years (depending on the type of data). Thus, current genomic research is characterized by extensive searching through databases for correlations with available data. In parallel, other steps of drug discovery have been automated, for example, automatic screening of drug candidates.

A current paradigm for drug research is as follows:

(a) identifying genes which meditate and/or are involved in diseases and/or biological processes of interest;

(b) establish identity and/or gene class of the identified genes;

(c) selecting genes with the most useful biological properties;

(d) screening for a lead compound which modulates the selected gene(s) or gene product(s);

(e) selecting related compounds which are even better; and

(f) testing the compound for specificity, toxicity, etc.

Unfortunately, this process fails to find a useful gene in over 90% of the cases. One challenge is to identify genes with useful biological properties and to limit that selection to genes which have a better than average chance of passing the testing stage.

Steps (a)-(c) are generally done on a basis of individual genes, by highly qualified workers. As the amount of available biological data increases, it is desirable to increase the integration of such data. However, the available data has several problems:

(i) Scale. There is an enormous amount of data. Even the number of different databases is large (>300). Also, the number of available tools is increasing, currently at over 100.

(ii) Update rate. The data is continuously being updated, so that a search which fails one day, might succeed if performed a short while later, as new data becomes available.

(iii) Complexity. Most of the data is not numerical. Instead the data is often string data, encoding relationships such as "member of" "property of" and "variant of".

(iv) Heterogeneity. The same piece of data may be described in different ways in different databases. In addition, each one of the databases uses an different format and has a different specialty.

(v) Garbage. A significant amount of the available data is simply garbage, artifacts and erroneous conclusions. Often, two databases contain conflicting data.

Bioinformatics, as a discipline, provides tools for researchers to manage and/or analyze the large amounts of data. A differentiation should be made between data and information. In "The Information Age", W.S. Davis and A. McCormake distinguish as follows: "Data are facts; information is the meaning which human beings assign to these facts. Individual elements of data, by themselves, have little meaning; it's only when these facts are in some way put together or processed that the meaning becomes clear". Thus, the role of a researcher in bioinformatics is to convert huge amounts of biological data into knowledge. Knowledge is sometimes defined as a set of conclusions and inferences implicated by the accumulated information. Thus, knowledge promotes an understanding that can lead to decisions and actions.

The concept of "intelligent agents" has arisen in the field of artificial intelligence, as describing software units which are capable of independent action and interaction with a surrounding. Independent action means that the agents are supplied with a facility for reasoning about the surrounding.

One type of reasoning facility is an inference engine, which contains rules about the interpretation of- and action on- facts sensed from the surroundings. Generally, when a

situation matches a certain rule, that rule fires, the result of which is the performance of the action.

International Business Machines Corporation, of Armonk, N.Y., is developing "Agent Builder" a commercial system for creating intelligent agents, in which each agent includes a reasoning engine based on inference rules.

SUMMARY OF THE INVENTION

The invention described herein, in some embodiments thereof, allows a fundamental change in the way genes are evaluated for the purposes of drug target discovery. Currently the dominate paradigm for gene evaluation requires that research analysts use one database or tool at a time, on one gene at a time and integrate and collate the resulting information before moving on to the next database, tool, and/or gene. Although there are extensive numbers of databases and tools at the disposal of biologists today, only a few are typically used because of the effort involved. After several such actions, the integrated data are correlated in order to detect implications in the data, and in order to prioritize the genes for further possible drug target consideration. Given the current explosion in genomics data (primarily genes), brought about by the emergence of high throughput genomics technologies, together with the requirement for exhaustive searching and analysis of the data before the very rare drug targets can be found, new methods for integrating the data presently available are probably required to assure the success of genomics-based drug discovery.

In accordance with a preferred embodiment of the invention, success in this genomics-driven drug development field is enabled by a break-through in the way this vast amount of complex data are accessed, integrated, and correlated to discover actionable conclusions (knowledge). Unfortunately the most popular 'data mining' and other tools available in other problem domains are often not useful in the genomics context, because of the scale, complexity, and heterogeneity of the data. The present invention, in some preferred embodiments thereof, provides a novel and powerful solution to the problem in the form of an 'intelligent' agent architecture which is specifically designed and optimized for the nature of the problem and the associated data.

One object of some preferred embodiments of the invention is to provide an automated system which enables a biological researcher to deal with challenges caused by problematic data. Preferably, the system aids the researcher to overcome challenges caused by one or more of the following: scale, updating, complexity, heterogeneity and garbage. Preferably, the system relieves the researcher of a need to become an expert in manipulating the bewildering range of available data sources and tools.

Another object of some preferred embodiments of the invention is to coordinate information generation, acquisition, utilization and/or distribution, especially in an industrial genomics enterprise setting.

Another object of some preferred embodiments of the invention is to assist and/or replace certain activities of a human researcher, at least in a data integration and/or knowledge discovery stage.

Another object of some preferred embodiments of the invention is to integrate genomic information with application related information, preferably from heterogeneous sources. Preferably, a wide survey of information integrated from heterogeneous sources aids in the selection of candidates for further testing. Preferably, the applications include development of pharmaceuticals, cosmetics, food additives, pesticides, herbicides and other biological-acting materials. Preferably, the application related information comprises disease pathology and biological function. Preferably, a knowledge discovery system in accordance with a preferred embodiment of the invention maximizes knowledge discovery and increases the level of success in new drug target identification.

One aspect of some preferred embodiments of the invention relates to discovering data, information and knowledge about biological data, especially genomic data. In a preferred embodiment of the invention, a discovery process is applied to a plurality of databases and/or analytical tools, using gene tokens and/or their attributes as search keys. Preferably, a knowledge discovery system also includes an inference unit which applies inference rules on the gene tokens and/or on data arriving from external data sources. The following is a skeleton of an automated knowledge discovery process which may be used in some preferred embodiments of the invention:

(a) defining a topic of interest and generating a gene token (some representative form of a gene), alternatively the gene token is retrieved from a database of gene tokens;

(b) identify missing data and define a set of data finding (a.k.a. data mining) goals;

(c) formulating one or more queries based on the data mining goals;

(d) identify one or more relevant databases, data sources and/or data analysis tools;

(e) formulating appropriate queries for each data source, communicate the query to the data sources and execute them;

(f) examining the hit list (query response) to identify new hits;

(g) extracting data from relevant hits;

(h) assembling the extracted data from all the data sources, yielding information;

(i) analyzing the integrated information and/or comparing it to other information, yielding inferences and conclusions, i.e. knowledge;

(j) facilitating conclusions and rational actions from the knowledge; and

(k) iteratively repeating the cycle (from step b). In some embodiments, at least some of the above steps may be varied, skipped and/or be performed in a changed order. In a preferred embodiment of the invention, the decision to skip or change an order of a step is made by a knowledge discovery system, preferably depending on circumstances as recognized by the system.

One aspect of some preferred embodiments of the invention is that the cycle is performed on one gene token at a time. A complete round of data mining and knowledge discovery cycles preferably includes applying the cycle once for each gene token in the gene token database. Thus, the state of knowledge for all the gene tokens in the gene token database is advanced substantially in parallel. Typically, one does not know at the start of the process which gene will yield a significant amount of knowledge.

An aspect of some embodiments of the present invention relates to goal setting in a knowledge discovery system. In a preferred embodiment of the invention, the system does not work towards a single predefined goal. Rather, a "target area" of desired results is defined and any hit within the target area is considered a success. Alternatively or additionally, any knowledge accumulated during the discovery process may be useful, for example for advancing the state of knowledge about particular genes. The target area is preferably not explicitly defined. Rather, useful properties of results falling in the target area are defined. Alternatively or additionally, the system encodes heuristic rules (preferably as inference rules) for advancing knowledge towards the desired target area.

An aspect of some preferred embodiments of the invention is that each gene token is associated with one or more schemes (or frames, slots or other similar AI constructs). Thus, missing and/or possibly available information can be determined from the scheme. In a preferred embodiment of the invention, flexible and/or extensible schemes/frames are used. Thus, widely varying attributes of gene tokens may be accommodated. Another advantage of some embodiments is that different gene tokens will have different extents of knowledge, and the schemes/frames can offer the necessary flexibility to accommodate this. Another advantage of some embodiments is that an automated scheme can be used to represent a significantly more complicated data structure that can be conveniently utilized by a human researcher, for example for setting data requirement goals and/or for analyzing the resulting data. Although a user may be able to enumerate all the possible slots and/or interactions between slots in a

scheme for a disease, humans are notoriously limited in their ability to mentally manipulate multiple-element/multi-relationship conceptual structures, as might be exemplified by interrelating enzymatic pathways and/or expression control mechanisms. A typical number quoted is that humans can manipulate 7±2 concepts. A computer program, on the other hand can easily manipulate a network of 10, 20, 50 or even 100 interrelated elements.

An aspect of some preferred embodiments of the invention relates to a continuously operating, possibly unattended, knowledge discovery cycle. In a preferred embodiment of the invention, gene tokens from the token database are continuously being evaluated for the purpose of knowledge acquisition. Thus, the system automatically takes into account any advances in the know-how and/or available resources, both internal to the system and external to the system. In a preferred embodiment of the invention, the control structure of the system is embodied, at least in part, as a set of inference rules, background assertions and/or knowledge representations.

An aspect of some preferred embodiments of the invention is self monitoring activity of a knowledge discovery system. In a preferred embodiment of the invention, interesting and/or erroneous activities are brought to the attention of relevant personnel, for further action by the personal. It should be appreciated that such self monitoring is especially important in a continuously operating system. In a preferred embodiment of the invention, self-monitoring includes monitoring data sources to determine which data sources are more dependable and/or which fields and/or subject matters thereof are more dependable. Alternatively or additionally, other aspects of data sources may be monitored, for example availability, response time and/or overlap with other databases. Alternatively or additionally, to monitoring databases, also tools which are activated for data search and/or analysis may be monitored.

Alternatively or additionally, self-monitoring includes monitoring the activity of the system itself. In one example, the system utilizes inference rules and self monitoring may include determining which rules are activated and/or their frequency of activation.

Alternatively or additionally, self-monitoring includes monitoring amounts of time spent on certain activities, such as knowledge discovery for certain gene tokens. Preferably, a user may define a limit on time spent on each gene token in general and/or on particular gene tokens. The system preferably includes a timer, to monitor time limits.

Alternatively or additionally, self-monitoring includes monitoring the progress of knowledge accumulation, for example as compared to predetermined standards and/or task and/or gene token specific standards sets by a user. Certain actions by the system, for example

stopping work on a gene token or reporting the gene token, may be triggered when the volume and/or quality and/or nature of the data for a particular gene token reach a particular state.

An aspect of some preferred embodiments of the invention relates to corrective action performed when an error is detected. Data errors may be detected by comparing data from different databases. Alternatively or additionally, the data in a database may not be consistent. Alternatively or additionally, the data in a database may be updated. In a preferred embodiment of the invention, when a knowledge discovery process is complete (or reaches a certain stage) the conclusions are compared to the data used to reach the conclusion. Some conclusions may trigger re-examination of the data, or correlation with other data, in order to detect inconsistencies or errors in the data. Although some inconsistencies are caused by erroneous data, some may be caused by a misinterpretation of available data, too broad generalizations, application of erroneous rules and/or erroneous application of rules. In a preferred embodiment of the invention, when an error is detected in a database, inferences based on the erroneous data may be deleted and/or reapplied. Alternatively or additionally, the data in the erroneous database is corrected, preferably with an indication in the database of the correction. In an external database, the system may generate a communication, such as an e-mail explaining the detected error and reasoning behind the error. When the data source is a programming tool, such an indication may be sent to a programmer and/or a maintainer of the program. When the data source is a laboratory, the error may be sent to laboratory personnel. In a preferred embodiment of the invention, the knowledge discovery system includes a plurality of rules which attempt to explain possible sources of data errors, based for example, on other available data.

An aspect of some preferred embodiments of the invention relates to automatic modification of system behavior, responsive, for example, to self-monitoring results. In one preferred embodiment of the invention, the system selects databases which have a lower determined error rate and/or higher availability for the searched for data. Alternatively or additionally, the system optimizes various parameters thereof, in response to the self monitoring. Alternatively or additionally, the system modifies rule results and/or rule activation, based on the self monitoring.

An aspect of some preferred embodiments of the invention relates to prioritization and/or resource allocation. In a preferred embodiment of the invention, various gene tokens are attributed a higher priority than other gene tokens. Knowledge discovery may be performed more rapidly, more often and/or utilizing more computationally expensive tools and/or resources, on tokens having higher priorities. In one example, human resources may be

allocated to higher priority tokens. Alternatively or additionally, more than one knowledge discovery cycle may be applied per round on the higher priority tokens. In a preferred embodiment of the invention, more than one type of priority is defined, relating, for example, to relative allocation of different resources.

An aspect of some preferred embodiments of the invention relates to overcoming errors in biological data by utilizing synergistic aspects of the data. Although biological data per-se often contains many erroneous and/or imprecise data elements, when the same data is found in different data sources, the reliability of the data is increased. In addition, the correlated data may include more breadth and/or depth of content. In a preferred embodiment of the invention, the transition from data to knowledge is dependent on the quantity and/or quality of supporting data. Thus, even if an inference could theoretically be based on a single element of data, the inference may require two or more supporting elements of data, possibly from unrelated databases and/or from unrelated data generation methods. Alternatively or additionally, weak inferences, such as those in which a reliability level is low may be marked and rechecked at a later time.

An aspect of some preferred embodiments of the invention relates to considering biological relevance against a plurality of dimensions. Preferably, the higher the number of matching dimensions, the higher the considered relevance. In a preferred embodiment of the invention, the dimensions include one or more of spatial, temporal, event and genome dimensions. The spatial dimension includes the spatial (in the human body) location of the manifestation of the relevant data. The temporal dimension includes an indication of a development stage and/or triggering conditions. The event dimension includes an indication of biological (result) events thought to be affected by the gene token. The genome dimension includes, for example genomic location, such as chromosome maps.

Another aspect of some preferred embodiment of the invention relates to logistical control of information, communications and workflow. In a preferred embodiment of the invention, the knowledge discovery system facilitates and/or controls the flow of information, for example, to assure that data is sent to interested parties and/or to avoid data overload at data recipients. Alternatively or additionally, the system suggests forging communication links between users who are not otherwise communicating or are not listed as communicating on certain subject matter. Alternatively or additionally, the system disseminates data, information and/or knowledge to recipients, who, in the system's judgment, would desire to receive that content.

In a preferred embodiment of the invention, the system automates workflow, for example, by generating work orders and/or prioritizing and/or allocating physical resources, such as laboratories, personnel and/or computer equipment. In a preferred embodiment of the invention, the system can directly or indirectly (through control programs) operate laboratory equipment, for data generation.

An aspect of some preferred embodiments of the invention relates to controlling data generation devices. In a preferred embodiment of the invention, the system may allow lower quality or reliability biological data to be provided. If later gleaned knowledge requires it, higher quality data may be requested by the system, for example via a work request sent to laboratory personnel. Alternatively or additionally, when requesting data the system defines a required quality level. Thus, the system can optimize throughput of laboratories, so that they operate at peak performance, generating data having only as high a quality or reliability as required.

An aspect of some preferred embodiments of the invention relates to automation level. Various levels of automation may be applied in different embodiments of the invention, at one end of an automation scale is a system which, once programmed, does not require any human input. On the other end of the scale is a system which will not operate unless it is continuously being controlled by a human. In between, are systems which report some or all activities, systems which request an OK for some or all activities, systems which notify a user of some or all happenings and/or systems which operate autonomously for a certain period of time, before requiring a user OK to continue. In a preferred embodiment of the invention, the system includes a plurality of automation levels. Preferably, different activities and/or results are handled differently. In one example, limits on resource expenditures may be defined. Going over the limits may require a user intervention. In another example, some results are determined to be important enough to be reported to a user immediately and some less urgent results are accumulated for a periodic report.

An aspect of some preferred embodiments of the invention relates to automatic notification of users, regarding newly available content (e.g., data, information and/or knowledge). An important consideration is not to overload a recipient with too much information. Additionally, information and/or knowledge should preferably be presented only when it's content and/or confidence level meet certain criteria. Additionally it should be noted that there is often no "requester" for information. Rather, the information and/or knowledge may have simply come up. In a preferred embodiment of the invention, a set of inference rules is defined to determine if data, information and/or knowledge should be present and who are

suitable recipients. In addition, when knowledge (implications, conclusions, correlations) are reported to personnel, links to the supporting data or links to visual presentations of the data which support the knowledge are preferably included. When a conclusion is reported, the system can preferably provide an 'audit trail' leading from the conclusion back to the data which support it. Preferably, when a user checks on a conclusion reached by the system, such checking includes an examination of the data used by the system, as such data is often erroneous.

An aspect of some preferred embodiments of the invention relates to basing an interaction with a user on known behavior, capabilities and/or desires of the user. In one example, data dissemination depends on a tolerance level of a user to information, on a workload and/or on a professed desire to acquire certain content and/or track a certain gene token. In another example, an explanation level and/or type of the system will be dependent on user characteristics. In another example, a work order provided to a user will be dependent on user characteristics.

An aspect of some preferred embodiments of the invention relates to encoding various biological heuristics in the knowledge discovery system. These encoded heuristics are preferably used to convert biological, genomic and/or drug target data into information and/or knowledge. In one example, properties may be assigned to gene tokens based on sequence identity or similarity, functional similarity, synteny and/or analogy of biological consequence. In particular, the heuristics may consider the form, participants, results side-effects, metabolic activity and/or interactions of pathways that maintain similarities across biological species and/or cell types.

An aspect of some preferred embodiments of the invention relates to various types of "gene" tokens. In a preferred embodiment of the invention, a gene token is used as a nucleation center to accumulate content about a particular gene. Alternatively or additionally, the token may be associated with a pathway and/or a family of genes. Alternatively or additionally, the token may be associated with a pathology, so that relevant genes are "discovered" and associated with the pathology.

An aspect of some preferred embodiments of the invention relates to multiple types of data associated with each gene token. Preferably, the data types include genomic and other biological data, prioritization of the token and possible interpretation of the data associated with the token. In a preferred embodiment of the invention, the system modifies data of one type in response to inference rules which are fired depending on data of a different type. In one

example, the prioritization of the token is modified responsive to an analysis of the biological data.

An aspect of some preferred embodiments of the invention relates to various methods of prioritization of gene tokens for usefulness in a particular application. Preferably, the tokens are ranked according based on biological data. Alternatively or additionally, other data pertaining to the application may be utilized. Preferably, the application for which the genes are ranked is a drug discovery application. One method is based on the biological relevance of gene tokens, for example, matching across biological spatial/temporal/event/genomic map dimensions, described above. Another method is based on pharmaceutical tractability. Some types of proteins may be more difficult to design drugs for; for example due to functional and/or structural characteristics and/or due to a difficulty in properly localizing the drug. Alternatively or additionally, some proteins may be similar to housekeeping proteins and/or significant other proteins, increasing the possibility of side effects. Alternatively or additionally, some proteins may be less suitable targets due to their expression profile and/or existence of parallel pathways, which operate even if the protein is knocked out. Another method is based on experimental tractability. As can be expected, experimental validation of a drug target is simpler if the gene has a close homologue in a different animals (for example a mouse). Also, proteins which cross the cell membrane may be easier to work with than proteins which remain in the cell nucleus. In a preferred embodiment of the invention, the above considerations are encoded as inference rules, background assertions, and/or criteria.

An aspect of some preferred embodiments of the invention is the utilization of knowledge discovery cycles including a plurality of points at which inferences may be applied. Generally, different inference rules are applied at each point. Preferably, the points at which inference rules may be applied include one or more of: data retrieval goal definition, data source selection, search results analysis and extraction, data integration, reporting/workflow to users and prioritization of gene tokens.

An aspect of some preferred embodiments of the invention is that inferences are preferably based on a meaning of terms (semantics), rather than on an exact word match. In a preferred embodiment of the invention, terms are interpreted using a semantic network, whose content is preferably derived, at least initially, from UMLS (unified medical language system). Thus, the terms IL-2 and interleukin 2 will be interpreted as the same concept. Alternatively or additionally, the broadening of a term may be applied at a query construction stage ('query expansion'). Preferably, the broadening for a particular database is limited to exclude terms

which are known not to be in use in the database and/or exclude terms which occur far too frequently to be meaningful.

An aspect of some preferred embodiments of the invention relates to a resource adapter hierarchy for accessing external data sources. In a preferred embodiment of the invention, the knowledge discovery system comprises one or more individual agents. Each such agent, preferably independently, communicates with data sources via adapters. The number of available data sources is numbered in the several hundreds, and the number is growing rapidly. In a preferred embodiment of the invention, the adapters are contained in a program interpreter (called the 'interface layer') which is embedded in the agent backbone. Adapters are constructed using tools which ease maintenance and/or creation of new adapters. One such tool is that adapters are preferably written in a text-processing language, preferably an interpreted language, for example Perl. Preferably, the language used supports complex representations of data, for example including object-oriented classes, to match the complexity and heterogeneity of the data sources. An advantage of using an interpreted or scripted adapter, in some preferred embodiments of the invention is rapid development, modification, and debugging. Thus enabling the hundreds of data sources to be accommodated in reasonable time. Conversely, the use of a compiled language might, in some cases, slow down development to the point that insufficient numbers of data sources could be accommodated to support effective knowledge discovery. The fragmentary and largely incomplete data for any one particular gene, means that the chance of serendipity (knowledge discovery from chance correlations) will probably be maximized only by accommodating as many adapters (i.e. data sources and tools) as possible. Indeed, successful drug target discovery may require this. In some embodiments of the invention both compiled and interpreted adapters may coexist in a single agent.

Alternatively or additionally, the adapters are organized into families of adapters, to utilize similarities between data sources. Preferably, the organization is hierarchical or hierarchical like, preferably including inheritance between adapters. Preferably, the hierarchy includes at least two, preferably at least three hierarchical levels. In a preferred embodiment of the invention, adapters can be complex objects which contain ('aggregate') other re-usable objects (such as 'protocol' objects), which preferably contributes to rapid development of adapters.

Alternatively or additionally, the adapters are defined with parameters, which may be modified, to exactly match an adapter to a particular database. Alternatively or additionally, the adapters are programmed in a modular manner, so that components of the adapter (e.g. specifying internal vs. external copies of a data source) may be replaced, if necessary. This

preferably enables an adapter to be rapidly reconfigured to begin working with an internal copy of a database that was once accessed externally, for example over the Internet and/or a dial-up connection.

An aspect of some preferred embodiments of the invention is that an adapter registers itself, when it becomes available to the agent. Thus, inference rules for selecting a relevant data source have available a list of available adapters, access properties (such as time and cost), field list, range of available data and/or other information which may aid in selecting a data source for retrieving data.

An aspect of some preferred embodiments of the invention relates to communication between adapters and an agent. Preferably, the adapters each communicate with the inference part of the agent using their own mailbox. Alternatively or additionally, a different mailbox is provided for each type and/or priority of message. Alternatively or additionally, a single mailbox is shared among some or all the adapters. In a preferred embodiment of the invention, each adapter operates as a separate computational thread, thus, the agent is not required to wait for requested information to be retrieved. Preferably, the adapters can communicate between themselves, for example, to coordinate usage of communication resources. Alternatively or additionally, such communication is performed through the inference engine.

In some cases, the agent may not wait for any information. Thus, some cycles might be devoid of retrieving results and contain only information analysis. Other cycles may include only data requesting and no analysis, as no new data is available. Preferably, the agent does pause for data retrieval at least a limited amount of time. Alternatively or additionally, the agent pauses responsive to a number, content and/or time stamp of outstanding data requests. Alternatively or additionally, an agent may pause until a particular data element is retrieved. Preferably, such pausing is dictated by suitable inference rules. Alternatively or additionally, an agent may select a faster- or a slower- accessed version of database, for example if a same database is accessible both on a local drive and externally. Such a selection may be responsive to an update status of the database and the instant importance of receiving up-to-date data. Alternatively or additionally, such a selection may be responsive to communication bottle necks. In one example, the inference engine may provide an adapter with a set of data sources, data from any of which may be sufficient. In another example, an agent may request data from a lower quality source, to be received faster and, in parallel, request data from a higher quality source (for example an on-line mirror of a local database), to be used to possibly correct inferences, when the data arrives.

An aspect of some preferred embodiments of the invention relates to activating several agents in parallel on a single gene token database. In a preferred embodiment of the invention, each agent includes a different set of inference rules and/or parameter settings. In some embodiments, knowledge determined by a first agent will not be available to a second agent, until it is released by the first agent. Alternatively or additionally, a total set of expertise and capabilities may be partitioned among a number of agents, and data retrieval and knowledge discovery tasks can be carried out through collaborations of specialized agents.

There is therefore provided in accordance with a preferred embodiment of the invention, a method of genomic data discovery, comprising:

(a) providing a gene data base comprising at least 10 genes;

(b) selecting one of said at least 10 genes;

(c) discovering knowledge for said selected gene;

(d) repeating said (b) and (c) for a plurality of said genes; and

(e) repeating said (b)-(d) a plurality of times such that knowledge is discovered substantially in parallel for all the selected genes.

Preferably, said (b)-(e) are performed substantially without human intervention. Alternatively or additionally, the method comprises automatically evaluating said genes for which knowledge has been discovered. Preferably, automatically evaluating said genes comprises ranking said genes according to their suitability for being drug leads. Alternatively or additionally, the method comprises deciding on further selecting of said genes in (b), responsive to said evaluation.

In a preferred embodiment of the invention, (c) comprises determining data needs for said genes. Preferably, each of said genes is associated with a scheme and wherein determining data needs comprise analyzing said scheme to determine data needs. Alternatively or additionally, (c) comprises formulating queries to obtain said needed data. Alternatively or additionally, (c) comprises setting up sub-goals for obtaining said data needs.

Alternatively or additionally, the method comprises selecting suitable databases for said data needs. Preferably, selecting suitable databases comprises selecting at least 10 databases for at least 5 needed elements of data.

In a preferred embodiment of the invention, (c) comprises receiving data for a plurality of data sources. Preferably, the method comprises integrating said received data. Alternatively or additionally, the method comprises analyzing said received data to produce knowledge. Preferably, the method comprises returning said selected gene to said database, in association with said discovered knowledge.

In a preferred embodiment of the invention, said at least 10 genes comprises at least 50 genes. Alternatively or additionally, said at least 10 genes comprises at least 100 genes. Alternatively or additionally, said at least 10 genes comprises at least 300 genes. Alternatively or additionally, said plurality of genes comprises at least 20% of said at least 10 genes. Alternatively or additionally, said plurality of genes comprises at least 50% of said at least 10 genes. Alternatively or additionally, said plurality of genes comprises at least 80% of said at least 10 genes.

There is also provided in accordance with a preferred embodiment of the invention, a method of genomic knowledge discovery, comprising:

determining at least one required data element for at least one gene;

querying a plurality of at least 50 databases for said at least one required data element;

receiving responses from said databases; and

analyzing said responses to increase knowledge for said at least one gene.

Preferably, said at least 50 databases comprise at least 100 databases. Alternatively or additionally, said at least 50 databases comprise at least 300 databases. Alternatively or additionally, said databases are all queried for a same data value. Alternatively or additionally, said method is performed automatically.

There is also provided in accordance with a preferred embodiment of the invention, a method of automated knowledge discovery, comprising:

continuously operating a knowledge discovery cycle comprising:

querying a database to receive data; and

performing inferences on said data to generate knowledge; and

re-evaluating said inferences when said database is modified

Preferably, said cycle is continuously operated over one week. Alternatively or additionally, said cycle is continuously operated over one month. Alternatively or additionally, said cycle is continuously operated over six months.

There is also provided in accordance with a preferred embodiment of the invention, a method of genomic knowledge discovery, comprising:

(a) selecting a gene token;

(b) determining data requirements for said gene token;

(c) requesting and receiving data responsive to said data requirements;

(d) analyzing said received information to increase knowledge for said gene token; and

(e) repeating (b)-(d) for the same gene token, at least 50 times.

Preferably, at least 50 times comprises at least 100 times. Alternatively or additionally, at least 50 times comprises at least 200 times.

There is also provided in accordance with a preferred embodiment of the invention, a knowledge discovers system comprising:

a first unit for determining data needs and analyzing data returned responsive to said needs; and

a plurality of at least 10 adapter units, for accessing at least 10 dissimilar data sources to provide said needed data.

Preferably, said first unit comprises an inference engine comprising rules for analyzing biological data. Preferably, said biological data comprises genomic data.

In a preferred embodiment of the invention, said at least 10 adapter units comprise at least 50 adapter units for 50 dissimilar data sources. Alternatively or additionally, said at least 10 adapter units comprise at least 100 adapter units for 100 dissimilar data sources. Alternatively or additionally, said at least 10 adapter units comprise at least 300 adapter units for 300 dissimilar data sources. Alternatively or additionally, said at least 10 data sources comprise at least 10 data analysis tools. Alternatively or additionally, said at least 10 data sources comprise at least 30 data analysis tools. Alternatively or additionally, said adapters are programmed in an interpreted text processing language. Preferably, said adapters are programmed in language including classes. Alternatively or additionally, said adapters are programmed in a Perl-like language.

In a preferred embodiment of the invention, the system comprises a central adapter registry, wherein, each of said adapter registers its availability in said central registry. Preferably, said registry is implemented as assertions. Alternatively or additionally, said adapters register their data provision capabilities in said registry.

In a preferred embodiment of the invention, said first unit analyses said data requirements to determine selected ones of said data sources to query.

There is also provided in accordance with a preferred embodiment of the invention, a method of ranking genes for an application, comprising:

providing a plurality of gene tokens;

automatically applying application-specific ranking rules to said gene tokens; and

ranking the gene tokens responsive to said application of rules.

Preferably, said plurality of gene tokens comprise at least 10 gene tokens. Alternatively or additionally, said plurality of gene tokens comprise at least 30 gene tokens. Alternatively or additionally, said plurality of gene tokens comprise at least 50 gene tokens.

Alternatively or additionally, said application comprises drug discovery. Preferably, said application specific rules comprise biological relevance rules. Preferably, said biological relevance rules comprise rules which match said gene tokens to a disease model across a plurality of biological dimensions.

Alternatively or additionally, said application specific rules comprise rules which determine experimental tractability. Alternatively or additionally, said application specific rules comprise rules which determine pharmaceutical tractability. Preferably, pharmaceutical tractability determination comprises an indication of ease of finding an effective pharmaceutical. Preferably, pharmaceutical tractability determination comprises an indication of ease of finding a pharmaceutical with a low level of side effects.

There is also provided in accordance with a preferred embodiment of the invention, a method of genomic information analysis, comprising:

providing a first model of a biological relationship which interrelates a first plurality of genes or proteins;

providing a second model of a biological relationship which interrelates a second plurality of genes or proteins; and

applying inference rules to said first and second models to infer missing information.

Preferably, said applying comprises determining data needs for at least one of said models, based on said applied inference rules. Alternatively or additionally, said biological relationships comprise enzymatic pathways. Alternatively or additionally, said biological relationships are in different species.

There is also provided in accordance with a preferred embodiment of the invention, a method of automated genomic knowledge discovery, comprising:

analyzing a gene token to determine required data;

selecting at least one suitable human expert; and

querying the at least one selected human expert for the required data.

There is also provided in accordance with a preferred embodiment of the invention, a method of automated genomic knowledge discovery, comprising:

analyzing a gene token to determine required data; and

automatically generating a work order to a laboratory to generate the required data.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be more clearly understood from the following detailed description of the preferred embodiments of the invention and from the attached drawings, in which:

Fig. 1 is a flowchart of a knowledge discovery process for genomic data, in accordance with a preferred embodiment of the invention;

Fig. 2 is a schematic illustration of data flow in the knowledge discovery process of Fig. 1;

Fig. 3 is a flowchart of a process of updating a gene token database in accordance with a preferred embodiment of the invention;

Fig. 4 is a schematic illustration of an integration of a knowledge discovery system in an industrial setting, in accordance with a preferred embodiment of the invention; and

Fig. 5 is a schematic block diagram of a knowledge discovery system useful for Fig. 4, in accordance with a preferred embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

KNOWLEDGE DISCOVERY

Fig. 1 is a flowchart of a knowledge discovery process for genomic data, in accordance with a preferred embodiment of the invention. Fig. 2 is a schematic illustration of data flow in the knowledge discovery process of Fig. 1. The following is a description of an exemplary knowledge discovery process. However, it will be appreciated that the process may be varied in many ways without changing the spirit of the process.

Referring to Fig. 1, in step 20, a seed is provided, which will drive the knowledge discovery cycle. In a preferred embodiment of the invention, the seed and/or content associated with the seed (data, information and/or knowledge) are analyzed to determine content which is missing for further increasing the knowledge of the seed. In a preferred embodiment of the invention, the seed is fitted to a frame or a scheme (in the AI sense of the word). By analyzing the frame or scheme, missing information may be determined. In one example, a gene-frame may include a chromosome location of the gene. If the seed does not include that information, it may need to be obtained. In a preferred embodiment of the invention, a plurality of schemes are provided. Preferably, a single frame (or schema) is selected for the seed. Alternatively or additionally, a plurality of frames are selected and the seed is fitted to them in parallel. In some cases, data may be desired to bolster a low confidence level in an already stored item of data.

In step 24, queries are formulated to retrieve the missing content. In some cases, the data may not be directly available from any data source. In a preferred embodiment of the invention, intermediate data retrieval goals may be set by the system, so that once those goals are met (at a later cycle), the required data may be identified and/or retrieved. Preferably the system includes a set of inference rules for generating these intermediate goals. Alternatively or additionally, these intermediate goals are asserting by modifying the scheme of the seed.

In a preferred embodiment of the invention, a gene token may have associated with it an alert request, so that if data becomes available, it is forwarded to the gene token. Automatic notification is a feature provided in many databases and is usually based on a query which is re-run periodically at the database. If any new data becomes available, the user of the database (in this case preferably the knowledge discovery system), is notified. In one example, a gene token may "set up" an alert on sequence information for the gene. Alternatively or additionally to such an alert function being performed by outside data sources, the knowledge discovery system can also be used to periodically search relevant data sources and automatically notify users (or hibernating tokens, on which no action is to be performed pending new information) of newly available data.

In step 26, data sources which can respond to the queries are identified. In a preferred embodiment of the invention, each available data source is registered in a central registry. Preferably, the same data is queried from a plurality of data sources, to overcome problems caused by erroneous data, missing data and/or slow response times.

In step 28, the selected data sources are queried. In a preferred embodiment of the invention, the queries are adapted to match the particular conventions and/or formats of the selected data sources. In a preferred embodiment of the invention, the queries are adapted by specifying the data needs in terms of canonical concepts and translating them into terms and a syntax which are understood by the data source. This syntactic and semantic translation of the data needs (data mining goals) into queries is preferably carried out by each data source adapter. Alternatively, if the adapters are hierarchically organized, at least some of the translation may be performed a single time for a set of related adapters.

In step 30, the results are retrieved from the data sources. In a preferred embodiment of the invention, the results are parsed in order to determine their content. Some data sources return results in a relational format. Others however, return strings, for example " gene RAS was found in 20% of tissue type X". Such a string is preferably parsed and the term "found in ... tissue type" interpreted to mean the same as Occurrence(RAS, X)=20% and "20%" (where the query was "in what percentage of tissue X is RAS found"). In addition, some formats are standard in the field of genomics, for example pathways are usually written as "XXX pathway". In a preferred embodiment of the invention, if the results are ambiguous, semantic mapping techniques, described below are used. Alternatively or additionally, the content of a result may be used to constrain a description of a result. For example, a field value which is formatted similar to a radiation hybrid map location may be used to interpret a field name "map location" as "radiation hybrid map location".

In a preferred embodiment of the invention, the knowledge discovery cycle is not delayed for data which has not been received from a data source. thus, data requested for a particular cycle may become available only at a later cycle, at a later date. The system preferably integrates what it can this time around, and then gathers (or receives) additional information in a later data cycle. In a preferred embodiment of the invention, late arriving data may be disposed of and/or used as it arrives. Alternatively, the system may wait, at least for particular data queries. Alternatively or additionally, the system may send a "cancel query" message to a slow responding database.

In step 32 the results from the many data sources are integrated. If the same data is found in several data sources, the confidence level of that data may increase. This is true especially if the origin of the data is by different methodologies or different researchers. The integration of the data may be used to fill in missing "slots" in the frame of the seed. In a preferred embodiment of the invention, if the data retrieval failed (e.g., no response, empty response, too long a response, erroneous response or irrelevant response), steps 22-30 may be repeated, at least for certain required data or for certain data sources.

In step 34, the information/data associated with a particular seed is analyzed. As a result of the analysis, additional "slots" of the seed's frame may be filled. Alternatively or additionally, other types of knowledge may be created, for example, a determination that a particular gene is a viable drug target, or conversely, clearly not a viable drug target. In another example, a new frame may be provided for the seed, based on the acquired knowledge.

In an optional; step 36, a knowledge discovery system may act on the acquired knowledge, for example to inform interested parties. In any case, the cycle is preferably repeated, with the enhanced seed being provided as a seed for a next knowledge discovery cycle.

Referring back to Fig. 2, which show the data flow of a knowledge discovery process, seeds are preferably gene tokens from a gene token database 40. Gene tokens, which are preferably nucleation centers around which discovered knowledge may be arranged, are described in greater detail below, under a separate heading. A single token 42 is selected and converted to a token with data requirements 44. Token 44 is converted to a token with queries 46. Data from databases 48 is used to convert this token into a token with query results 50. Data integration converts token 50 into a token with information 52. Analysis of token 52 generates a token with new knowledge 54, which may be returned to token database 40 and/or forwarded to recipients 56. In a preferred embodiment of the invention, some or all of the above transformations and decisions are orchestrated and/or performed by rules from an

inference rule set 58. Inference rules are described in greater detail in a separate section, below.

In a preferred embodiment of the invention, database 40 itself is also one of the databases queried in step 28 of Fig. 1, since two gene tokens may refer to similar or overlapping things.

In the following example, a candidate GL375 is the seed. An example of data which is in the scheme is an association with a sequence EST 9224. Querying for a sequence matching the EST might yield that there is a match with osteoblast-specific GPCR and a map location of the gene. A second round of queries (at a later time) might yield the fact that Osteoporosis has been genetically linked with that same map location. An example of knowledge generation is that GL375 is a good candidate to account for osteoporosis.

In Fig. 1, not all the steps are necessary in order to increase the level of knowledge for each gene token. Additionally or alternatively, the order of steps may be modified. In one example, identification of suitable data sources is not performed. As a result, many databases may return empty responses. In another example, the information from the data sources is not integrated. Instead, the first arriving data or the data with the highest reliability value (based on the database identification and/or provided by the database) are used. In another example, data needs may be defined after the information is analyzed, so that a token is always associated with outstanding data needs.

Fig. 3 is a flowchart of a process of updating a gene token database in accordance with a preferred embodiment of the invention. Preferably, the knowledge discovery cycle of Figs. 1 and 2 are continually being performed. In Fig. 3, a round comprises performing one knowledge-increasing step for each gene token in database 40. Thus, the level of knowledge for the entire set of gene tokens is always advancing. As can be appreciated, it is not usually known in advance which of the gene tokens will yield a viable drug target.

In a preferred embodiment of the invention, the rounds are driven by selection of gene tokens from database 40. Alternatively or additionally, the rounds may be driven by external events, for example, an updating of data or a specific user request.

Fig. 4 is a schematic illustration of an integration of a knowledge discovery system 70 in an industrial setting, in accordance with a preferred embodiment of the invention. One preferred industrial setting is gene discovery and drug lead generation, in which genes are selected based on their suitability to become drug targets.

As shown in the Fig., system 70 may be connected to management 70, one or more external data sources 74, one or more internal data sources 75, one or more laboratories 76, a

first researcher 78 and one or more optional additional researchers 80, one or more project groups 82, including a project head 83, one or more bioinformatics tools 84 and one or more programmer/maintainer 86. Bioinformatics tools 84 may serve as data sources and/or tools used to analyze existing information. In some preferred embodiments of the invention, system 70 does not need to be connected to all of the above locations.

In Fig. 4, knowledge discovery system 70 performs two functions, knowledge discovery and knowledge management. Additional possible functions are described elsewhere herein. For knowledge discovery, system 70 is preferably connected to data sources 74 and 75 from which it retrieves data. Additionally, seeds (gene tokens) are preferably provided by researchers 78 and/or by laboratories 76. Results may be reported to researcher, project groups and/or management.

In a preferred embodiment of the invention, laboratories 76 comprise two types of laboratories, bulk laboratories, which generate a large amount of data, for example, ESTs, relative expression level data, and protein-compound affinity, and detail laboratories, which are designed to answer specific in-depth questions, for example, what tissues are affected when a particular pharmaceutical is ingested.

In a preferred embodiment of the invention, genomic data from the bulk laboratories is automatically input into system 70. In a preferred embodiment of the invention, not all genomic data generates a token. Preferably, system 70 includes a filter program which determines which of the received bulk data elements should be converted into a gene token and which not. Preferably, such a filter includes a matching unit which attempts to match new data elements with existing gene tokens. Alternatively or additionally, the determination may be based on characteristics of the data, for example, length of sequence and confidence level.

Referring back to the knowledge management function, it is noted that database 40 accumulates information which is of interest to many parts of an organization which includes system 70, not necessarily those parts which initiated the knowledge discovery or which provided the information. In one example, researcher 78 will initiate a gene token, but the results are of greater interest to researcher 80. In another example, work being performed in a laboratory 76 happens to be relevant to a project group 82. In another example, a management member 72 may be interested in a high level overview of the current state of knowledge in the organization. In another example, knowledge and/or information in database 40 might belong to and/or modify an internal or an external data source.

In a preferred embodiment of the invention, system 70 distributes information taking into account functional and/or personal characteristics of the recipient of the information

and/or taking into account special requests. Functional characteristics include, for example, a managerial level, a person being a head of a project (responsibility level), secrecy considerations (compartmentalization), and/or a person being part of a project, part of a research group or being a more generally open fielded individual. Thus, an interesting discovery regarding an osteoporosis associated gene may be communicated to a project working on drug leads for osteoporosis, to a researcher who generated a significant item of the information leading to the discovery, to a researcher in the field of body fluid calcium levels, to a manager (based on the level of interest of the item) and/or to a worker who requested to be kept up to date on osteoporosis related information. Personal characteristics include, for example, an annoyance level, number of items of information to be sent per week, a known workload, a personal interest field and/or interest level threshold.

In a preferred embodiment of the invention, information is sent by e-mail preferably including hypertext/HTML email with links to richer or interactive communications. Alternatively or additionally, the information is published on a network, for example using an Intranet or an Internet. Alternatively or additionally, the information is sent by fax. Alternatively or additionally, the information is sent by voice, for example directly into voice mailboxes of recipient.

In one embodiment of the invention, system 70 is formed of two sub-systems, a knowledge management system and a knowledge discovery system which is treated by the knowledge management system as one of its users.

GENERAL SYSTEM STRUCTURE

Fig. 5 is a schematic block diagram of a knowledge discovery system 70, useful for Fig. 4, in accordance with a preferred embodiment of the invention. the functionality of system 70 is preferably provided by one or more agent instances 102 which perform the above described knowledge discovery process. Each agent preferably comprises an inference engine 104, which is used to manipulate knowledge representations and/or assertions pertaining to gene tokens and/or diseases, and one or more adapter 110, which are used to transfer data to and from the agent and/or to otherwise communicate with an external world, for example, to control one or more tools 116. The collection of adapters are preferably contained in an interface layer 109, which preferably comprises a program interpreter (preferably Perl), which is optimized for text processing, inter-process communication and/or rapid development. Inference engine 104 is preferably written in CLIPS. An agent backbone 106 is preferably provided to link inference engine 104 with adapters 110 and to provide an operational setting for agent 102. Preferably, backbone 106 includes a mailbox 108, for adapters 110 informing

inference engine 104 of their activities and/or new data. Each adapter preferably has its own mailbox. Alternatively or additionally, each priority level and/or message type has its own mailbox, possibly shared between adapters and/or adapter hierarchy groups.

In a preferred embodiment of the invention, inference engine 104 is used for one or more of: setting data requirement goals, formulating queries, selecting relevant data sources, parsing query results, integrating data, analyzing information and/or deciding if and to whom messages are sent. Additionally, inference engine 104 may perform other activities described herein, for example prioritization of gene tokens. In a preferred embodiment of the invention, inference engine 104 is also used to control program flow, data flow and/or the overall behavior of system 70.

One advantage which may be realized using inference rules for controlling data flow is an increased ease in modeling real-life decision making strategies and rules used by bioinformatics researchers. Another advantage which may be realized is the provision of a more flexible control structure. Another advantage which may be realized is that a rule-based control structure is more data oriented than a procedural based control structure.

Alternatively, at least some of the control structures of agent 102 are provided by procedural or other methods. In a preferred embodiment of the invention, inference engine 104 is used to model particular control sub-structures. For example, re-applying a BLAST program on a data element, at varying levels of stringency (possibly at consecutive cycles), until a match is found.

In a preferred embodiment of the invention, all the contact of agent 102 with the outside world is through adapters, possibly including also file access. Thus, in Fig. 5, all the connections to elements outside agent 102 are through adapters. In a preferred embodiment of the invention, each adapter is used for connecting to a single data source or family of data sources. Preferably, when an adapter is activated, the adapter registers its existence in a data source registry 112, which may take the form of knowledge representations in the inference engine. In a preferred embodiment of the invention, data source registry 112 contains information useful for selecting data sources (step 26, Fig. 1) and/or for communicating with data sources, including one or more of:

(a) list of fields and their meaning and/or semantic mappings;

(b) communication protocol;

(c) required query format;

(d) output format;

(e) terms used in output (e.g. "is a variant of");

(f) terms useful in queries;

(g) formatting of particular types of values;

(h) ranges of values in fields and/or coverage of a database;

(i) a confidence level (per database or per different parts of a single database);

(j) expected response time; and/or

(k) update status.

In a preferred embodiment of the invention, data source registry 112 is embodied, at least in part, as assertions. Alternatively, data source registry 112 is provided by a mediator software component, possibly an adapter, which provides a list of relevant databases, when queried. In a preferred embodiment of the invention, the information in data source registry 112 also affects query formulation, for example by an adapter providing rules or assertions which indicate what data could be requested. Alternatively or additionally, such inference rules may be used to expand a scheme associated with a gene token. For example, a gene token scheme may include only sequence information. If an adapter is provided which matches sequences with protein acidity sensitivity, a new slot entitled "protein acidity sensitivity" may be added to the scheme. Such an event is preferably also indicated to programmer 86 (Fig 4), so that appropriate information analysis rules may be added.

It should be appreciated that a suitable adapter can communicate with any type of information source, including data sources, e-mail and/or data manipulation tools. In one example, an adapter may be used to run a BLAST homology test. In another example, a different adapter may be used to perform some or all of the information analysis step. Alternatively, an adapter may be used to execute a program which performs any of the steps described with reference to Fig. 1. In another example, a special adapter may be used to match schemes to existing gene tokens. Thus an adapter may send and/or receive various types of data, including data objects of agent 102, such as schemes or rules.

In a preferred embodiment of the invention, adapters are written in a text-processing language. Preferably the language is interpreted. Preferably, the language is Perl. In a preferred embodiment of the invention, adapter design takes into account similarities between data sources. In one example, the adapters are designed in a modular manner, so that only those modules which are different need to be programmed. Additionally, such modular design aids in updating adapters when data source formats change. Additionally, such modular design aids when a database is to be accessed in a plurality of methods, for example over the Internet and from a local CDROM drive.

In a preferred embodiment of the invention, the adapters are organized in a hierarchical structure ("object oriented"). Preferably, one type of adapter may inherit properties and/or procedures from a second adapter. In a preferred embodiment of the invention, a three or more level hierarchy is provided, for example: adapters||service_adapters||e-mail_adapter or adapters||data_source_adapters||Genebank_Family||Genebank1. Thus, a plurality of adapters which all access data sources having a format similar to Genebank, may be grouped together.

In a preferred embodiment of the invention, especially when the adapters are written in Perl or a similar language, the adapter programming utilizes the object-oriented capabilities of Perl to successfully communicate with data sources having widely varying formats and relationships between data elements. Alternatively or additionally, the data source adapters make use of a parser object (preferably written in Perl) which reads into memory a file containing parsing, syntactic transformations and/or semantic mappings necessary to process records returned from a particular data source. Alternatively or additionally, the adapters may be modified and or replaced while system 70 is executing.

In a preferred embodiment of the invention, at least some of the adapters utilize a standardized data access mechanism, for example SQL or OPM. These mechanisms are preferably implemented as re-usable software components (preferably Perl objects), which are preferably used in adapters, to speed their development.

The above design methods of adapters preferably enable rapid development of new adapters and/or maintenance of existing adapters, in the face of the large amount of heterogeneity found in bioinformatics databases.

One important different between data sources is that they may use different semantics to describe a similar concept and/or data value. There may also be some variability in a single data source, especially if it is not structured. However, many data sources only support key-word and/or field name based search. Thus, a search for IL-2 in one database may not yield proper results if the term used in the database is interleukin-2. In a preferred embodiment of the invention, semantic mapping is used to over come this problem. The semantic mapping may be applied during query formulation, to define a larger set of related key words. In one example, semantic mapping may be used to broaden a term, for example, heart tissue should include myocardium, papillary muscle, etc. Alternatively or additionally, semantic mapping may be used when broadening a query, for example by suggesting higher levels of abstractions. In one example, "intestinal muscle" may be broadened to "smooth muscle". Alternatively or additionally, the semantic mapping may be applied when adapting the query to a particular database. Alternatively or additionally, the semantic mapping is applied while

parsing the results of a query, so that information contained in the database's records are returned to the inference engine in a canonical or standardized vocabulary. The semantic mapping may be performed by the adapter for the data source. Alternatively or additionally, the semantic mapping is performed by the inference engine.

In a preferred embodiment of the invention, the semantic mapping is performed using a comprehensive database of biomedical terms 114, for example, initially populated with content from the Unified Medical Language System (UMLS) knowledge base, available from the National Library of Medicine. Preferably, the database is converted into inference rules and/or assertions. Preferably database 114 is continuously updated, for example by users or by system 70 itself. In one example, system 70 updates semantic mappings based on a query result in which such mappings are explicitly stated, for example in a text based database: "...cardiac muscle (e.g., papillary muscle)...", may be interpreted by system 70 to mean that papillary muscle is a type of cardiac muscle, at least in that database.

In some cases, a syntactic analysis may aid in the semantic analysis. For example as described herein with respect to radiation hybrid map locations, a format used in a field may be used to guess at or narrow the range of possibility to a semantic meaning of a field name. For example, a field that is labeled 'Map Location' (of which there are several possible types) and containing a value of '21cR' may be mapped to the concept of 'rad-hybrid-map-location = 21'.

In a preferred embodiment of the invention, system 70 is implemented on a LINUX machine with at least 64Mb. Preferably, a plurality of agents are simultaneously executed on a single machine. Alternatively or additionally, a plurality of agents are executed on a plurality of networked machines. In a preferred embodiment of the invention, the machine and/or the network are selected with redundancy in mind, so that system 70 will be able to continue normal operation even if a hardware or software component fails. One advantage of LINUX computational platforms is their excellent performance/price ratio. Thus, large numbers of agent instances may be effectively deployed to analyze a large number of gene tokens.

In a preferred embodiment of the invention, each agent is executed as a multi-threaded application. Preferably, each adapter is in a separate execution thread, as are the backbone, interface layer, and/or the inference engine. Hence, if one adapter is stuck, the rest of agent 102 is not significantly affected, leading to significant robustness. Alternatively or additionally, inference engine 104 may also be multi-threaded, especially to support multiple simultaneous inference chains. Also, it should be noted that an inference rule may depend on the execution of an adapter. Alternatively or additionally, the agent may be implemented in a multi-process

environment, for example with some of the adapters being separate processes. Alternatively or additionally, a client-server paradigm may be used, for example, with the inference engine being a "client" and the adapters being "servers". In one example, a single adapter may be shared among several agents and/or instances of inference engines. Alternatively or additionally, the agent may be executed in a network environment, for example with certain adapters being executed on separate machines and/or utilizing a distributed inference engine. Alternatively or additionally, an agent may comprise a plurality of sub agents, possibly programmed using dissimilar techniques (e.g., procedural, inference rules, neural network and pattern matching). In one example the inference engine comprises a data-source-selector sub-agent and a drug-lead-evaluation sub-agent.

GENE TOKENS

In a preferred embodiment of the invention, the seeds of Fig. 1 are gene tokens. Preferably, each gene token is a collection of data centered around a particular gene, which servers as a nucleation center. In a preferred embodiment of the invention, each gene token is associated with one or more frames (or schemes), into which the collected data is placed. The slots may be fixed slots, which allow only a fixed number of associated data elements. Alternatively or additionally, at least some of the slots may be flexible slots which allow any number of data elements. Alternatively or additionally, some of the slots may accept data only if it matches a certain desired condition of data or if the data has a very high confidence level. For example, a sequence slot may require the sequence to be over 100 bases long, since most genes are at least that long. Preferably, each token also maintains a pointer to and/or a copy of data and/or rules and/or queries used to modify the token.

Alternatively to a token being limited to being a gene (in a genomic data base), a token may be defined to match a group of genes or even if no gene is known. For example, a token may be defined for "the gene, a mutation in which, causes Alzheimer's disease". In another example, the token can be broadened to include a concept beyond a gene, for example a token may be a disease token, which type of token is used to accumulate knowledge about a disease.

In a preferred embodiment of the invention, gene tokens are created by users, preferably at a start of a task, but possibly also while the task is active. Alternatively or additionally, gene tokens may be created as a result of a query and/or as a result of an analysis of data. In the first case, a query may return two seemingly unrelated genes. One solution is to dispose of one of the genes. Another solution is to create a new gene token for the extra gene. In the second case, an analysis of a gene token may indicate that there are actually two or more genes involved, for example due to an unexpected number of inconsistencies and/or a bi-

phasic distribution. Conversely, two gene tokens may be merged, if there is a high enough match between them. Alternatively or additionally, related genes may be clustered and/or otherwise associated, in that they receive similar treatment.

In a preferred embodiment of the invention, schemes are created by users. Alternatively or additionally, schemes may be updated automatically, for example in response to a query result in which data fields, which are not supported by the scheme, are retrieved or as a result of parsing a free-form query response, which response contains a relationship not covered by the existing scheme. The association of one or more particular schemes with a gene may be performed manually. Alternatively or additionally, system 70 may suggest and/or associate the schemes. Alternatively or additionally, system 70 may modify the schemes and/or modify the scheme associations, for example based on inference rules which analyze an existing state of knowledge about a gene token.

In a preferred embodiment of the invention, one token may message a second token, for example if a piece of data can only belong to one of the tokens. The result of such a message may be an indication in one token that a previously correct data is not considered to be suspect. Another possible result is the activation of a mediation process, in which differences between the two tokens are settled. Another type of message is when the tokens belong to the same family of tokens and one token comes up with information which may be of interest to a second token. Preferably, the tokens in a family are explicitly linked, for example as part of their scheme, so that not all actions at one token will trigger an action at a related token. The above messages are preferably implemented by setting attributes of schemes associated with gene tokens, which attributes may initiate the firing of "message handling" rules.

It should be noted that in system 70 it is expected that one gene token be "researched" based on knowledge advances in a second token. For example, an osteoporosis token may utilize knowledge gleaned while researching the maintenance of serum calcium level. Such chaining of inferences between two, three or more tokens may be performed automatically, as described herein, for example by treating the gene token data base as a data source for data retrieval. In addition to chaining, other "collaboration" configurations are expected, for example, two or more gene tokens co-advancing, each one building on the other; star configurations where a single gene token is fed from a plurality of gene tokens; token rings, where each token is dependent on a pervious token; and more complex configurations, for example as can be described using a directed graph or tree. In a preferred embodiment of the invention, such interactions are detected by system 70 (preferably using suitable inference

rules) and possibly announced to suitable users, such as researchers. Alternatively or additionally, when such interactions are detected, the order of evaluation of gene tokens may be modified, to optimize the rate of data accumulation, for example, by scheduling the related gene tokens to be processed at a same time (possibly on different machines), or in an order which matches their dependencies.

In a preferred embodiment of the invention, each token is associated with several distinct types of information. Preferably, inference rules can be used to modify one type of information in response to a second type. The information types preferably include two or more of:

(a) genomic data, such as sequence, tissue specificity, gene function, protein homologue, chromosome map location, expression level and/or interactions with other genes;

(b) application data, such as disease phase, condition trigger, associated pathology;

(c) relevance data, such as possible interpretation of the gene token;

(d) importance data, which include priority levels of the gene token; and/or

(e) applicability data, which includes data relating to the usefulness of the gene token for various applications.

INFERENCE RULES

In a preferred embodiment of the invention, an inference rule contains one or more patterns which are used to scan a fact pool for facts (assertions) which match: e.g., the pattern '(expressed-in ?gene ?tissue ?level)' will match an assertion of the form '(expressed-in GL3456 liver 0.56)'. However, other formats and types of inferences rules are known in the art and may be used in alternative embodiments of the invention or in conjunction with the above type of inference rule.

Although inference rules have many uses in system 70, one of the important uses is to respond to data in a gene token by performing actions, which generally affect the data in the gene token.

In a preferred embodiment of the invention, the data in a gene token are first converted to a 1st order predicate form before being processed by the inference engine. The inference rules are preferably match driven rules which are fired if their matching conditions are met.

In a preferred embodiment of the invention, generating a conclusion requires a minimal confidence level. Preferably, this means that plurality of data elements must support the same conclusion. If a required confidence level cannot be met, the rules may not fire or the conclusion may be unrecognized Alternatively, a low confidence conclusion may be generated.

In a preferred embodiment of the invention, system 70 includes several possible rule sets and background assertions (knowledge). When a user sets up a task, the user preferably selects one or more rule sets to use in the make up of an agent 102. In a preferred embodiment of the invention, some rule sets may be specialized per task or per scientific outlook. Alternatively or additionally, a rule set may be crafted and/or modified so that it is particular to a disease and/or a disease model. Alternatively or additionally, a rule set may include inference rules which fire on schemes matches. Alternatively or additionally, the scheme may be embodied, at least in part, as rules.

In a preferred embodiment of the invention, one of the first rule types applied to a gene token is used identifying the gene token. In one example, a gene token is provided as a EST. The applied rule preferably identifies that a complete sequence is missing and searches in relevant data bases for sequences with homologies to the EST. The series of data mining and knowledge discovery goals may follow a particular line of reasoning, or a particular hypothesis exploration, which may possibly be a characteristic of the agent, gene token, associated schemes and/or rule set.

In a preferred embodiment of the invention, inference rules are used to set priorities between gene tokens. As explained herein, priorities may be used, inter alia, for resource allocation and/or for selecting recipients for data. Some gene tokens may be marked as not being active or may be advanced only once every few rounds. Preferably, different resources may have different priorities associated with them, per gene token. For example, two gene tokens may have opposing time priority and money priority.

In a preferred embodiment of the invention, prioritization is based on a score. The determination may be made by comparing the score to a threshold or by comparing the score to scores of other gene tokens. In a preferred embodiment of the invention, an effective score also takes into account changes in score as a function of expenditure of time, money, cycles and/or other resources. Possibly, the ranking score for a particular gene token is based on multiple components. Preferably, gene tokens whose score does not increase, are provided with a reduced priority. Alternatively or additionally, gene tokens are removed from consideration if knowledge discovery may be terminated, for example, if they have reached a complete description, have been identified as useful candidates, have been identified as house-keeping genes and/or if no further data has been found.

In a preferred embodiment of the invention, a score is generated based on matching to ranking rules. For convenience, these rules may be divided into the following types:

(a) rules which determine biological relevance of a gene;

(b) rules which determine pharmaceutical tractability of a gene;

(c) rules which determine experimental tractability;

(d) rules which determine research tractability; and/or

(e) rules which determine financial tractability.

As an example of biological relevance, we refer back to the example with gene token GL375. This token may be analyzed with respect to several biological dimensions. Preferably, a score may be generated which indicates the number and/or extent of matching biological dimensions. These dimensions may also be used to define a biological context or profile of the gene token, which may be used, for example, for identification and/or matching. In a preferred embodiment of the invention, the dimensions used are:

(i) spatial, e.g., is the gene expressed in a tissue related to the desired disease;

(ii) genomic, e.g., is there a match between the gene token and expected genomic information;

(iii) temporal, e.g., do changes in expression level of the gene track changes in disease state; and

(iv) event, e.g. does the function of the gene token match an expected function.

In the osteoporosis example, candidate GL375 matched all four dimensions (it was expressed in the bone, had a matching sequence, increased when disease appeared and was related to a bone building/breakdown function). Alternatively or additionally, these dimensions may be used to drive data requirements, by seeking to increase the number of available and/or matching dimensions.

As an example of pharmaceutical tractability, it is noted that, historically, there has been a greater success in desirably and selectively affecting proteins from certain families having certain functional and/or structural characteristics, for example G-protein coupled receptors, ligands, enzymes and protein channels, while other proteins, such as transcription factors have been less successfully affected. This may be detected by checking the protein family of a gene token and/or by homology. In another example, there may be an identified lead compound which interacts with a homologous protein. In another example, genes where a large amount of side effects are expected are usually not a good drug target.

As an example of experimental tractability, genes which have a mouse homologue are usually easier to evaluate functionally, since a mouse in which the gene is knocked out can be developed. Other experimental techniques also require biological knowledge, for example, some expression assays require that the full length sequence be identified and cloned.

As an example of research tractability, some genes families have been more heavily researched. Thus, more relevant biological information may be expected to be available. The volume of possibly relevant information is often an important consideration in knowledge discovery.

As an example of financial tractability, the importance of developing a pharmaceutical is often a function of the market for the drug. If only a small market is available, the payoff may be too small to take a risk. The risk is preferably assessed using the other prioritization rules and/or by providing marketing information.

It should be noted that the above rules may also be used to select research areas in which human researchers have not mined out, since they were expected to be unproductive.

In a preferred embodiment of the invention, inference rules are used to draw analogies between biological situations, in order to point at missing data and/or offer explanations. In one example, two pathways may be compared. For evolutionary reasons, similar pathways often utilize homologous genes. Thus, if one pathway is known, it is possible to expect to find homologous genes in a similar pathway (in the same animal or even between species). This type of homology may also be used for a negative purpose, for example to determine an increased risk of side effects.

AUTOMATION LEVEL

System 70 may be operated at various levels of automation. Alternatively or additionally, different components of system 70 may be operated at different levels of automation. In general, automation levels vary between a completely manual system, when a person is required to perform all the activities and make all the decisions and a completely automatic system, where the system does as it will and does not even notify a person after the fact. In a preferred embodiment of the invention, system 70 operates at an intermediate level. For example, the system may require human intervention for some decision, perhaps only the final decision after much inferencing and decision making has already been performed by the system. Also, the system may notify a person of a made decision. Also, a person may affect the operation of the system directly, for example by command, or indirectly, for example by modifying a control-type inference rule or by setting a goal.

In various preferred embodiments of the invention, any of the above described activities may be performed manually, automatically or semi automatically, meaning that system 70 aids a person in performing the step. Generally, it is preferred to limit human intervention, at least to those activities which i) are too difficult to program a computer for, or ii) for an ultimate decision involving the commitment of large amounts of human and/or other

resources. However, it can be appreciated that as inference chains become longer and data sources less dependable, it may become more desirable to receive a human input. To facilitate the involvement of human decision making in the process, system 70 will preferably do as much reasoning as possible proactively, report its conclusions to experts, and then facilitate the manual review/approval process by providing links to the intermediate conclusions/decisions and/or supporting data behind those conclusions. In essence, providing an 'audit' trail from high level conclusions, to intermediate conclusions, all the way back to supporting raw experimental data. By combining intelligent automation with appropriate levels of human intervention, an optimal balance of both knowledge discovery throughput and reliability is preferably obtained. In addition, a roll-back capability will preferably enable a user to undo inferences which are unacceptable and/or to test alternatives. In a preferred embodiment of the invention, the automation level appropriate for tasks associated with a particular user is learned, by tracking the type, degree and amount of roll-back required by that user, preferably, by task type.

In a preferred embodiment of the invention, the control of system 70 is by agents 102. Preferably however, users, for example, those listed in Fig. 4 may also exercise control, for example:

(a) a breakpoint;

(b) a report at a certain inference, priority level and/or found data;

(c) changing priorities of gene tokens; and/or

(d) modifying rules.

Although system 70 has not been generally described as a goal-oriented system, in a preferred embodiment of the invention, a user may define a goal to be met and/or reported when it is met. In a preferred embodiment of the invention, system 70 will include backward-chaining inference rules for working back from the goal to required starting data. Alternatively or additionally, system 70 will include a planner, for example to define a query or execution plan consisting of more or more steps needed to reach a desired information/knowledge state.

In a preferred embodiment of the invention, each agent 102 may be associated therewith a particular automation level. Such a level may be preselected to match a particular task, gene token, set of databases and/or user preference. Alternatively or additionally, the automation level may change while agent 102 is executing, for example, in response to received data, changing priorities of a gene token and/or suitable inference rules. Under circumstances like these, the agent may contact a human expert of which it is aware and/or associated with, and request approval or clarification on an action it is about to take.

USER INTERFACE

One important aspect of user interface has already been mentioned, matching data sent to a user to the user's desire, ability and/or job function (responsibility). In a preferred embodiment of the invention, system 70 includes a user model, to better guess if, when and how content should be sent to a user. Preferably the user model is based on psychological characteristics of a user, possibly entered by a user himself. Alternatively or additionally, the user model is based on a task being performed by the user and/or on his responsibilities. Thus, new discoveries may be reported immediately to a researcher, while such new discoveries will be ranked and reported to a manger once a week. Additionally or alternatively, the model for a particular user could be learned during interactions with the user over time.

In a preferred embodiment of the invention, one or more of the following human interfaces are provided:

(a) e-mail, sending and receiving;

(b) WWW pages, for filling forms and/or for receiving and/or transmitting graphics-rich and/or interactive responses;

(c) command line prompt, for commands;

(d) menus, preferably for commonly used commands; and/or

(e) by receiving and/or generating electronic files.

In a preferred embodiment of the invention, the type of interaction can be immediate or less so, preferably depending on an urgency; for example:

(a) setting up tasks to be performed, for example gene tokens and/or inference rule sets, or changing inference rules, which actions do not generate a response in the near future;

(b) setting up breakpoints and/or report-points;

(c) requesting a status report and or a display of data base contents;

(d) scripts to be executed under certain circumstances;

(e) other ad hoc commands, for example requesting a roll-back of an inference chain; and/or

(f) real-time or semi real-time question/answer type interactions.

In a preferred embodiment of the invention, results are displayed as a ranked list. Preferably, a graphical interface is provided for a user to examine the results, preferably using a point-and-click interface. Such examination may include displaying an inference chain, displaying activated rules, displaying confidence levels, displaying raw data, displaying data sources and/or displaying inconsistencies, within the gene token and/or with other sources of data. Preferably, the data is displayed in a hierarchical manner. for example, clicking on a gene

token will display statistics of confidence levels and rule applications. Clicking on rule applications will display a chain (tree) of applied rules. Clicking on a rule will display data used to match the rule. Clicking on a data element will display the source of the data. Preferably, a user may annotate any viewed item, for example for later viewing or for a different user, such as an operator. For example, an operator may view all the annotations associated with a particular rule, to better analyze problems with the rule. In some cases, a single gene token is analyzed at a time. In other case, a plurality of gene tokens may be viewed and/or analyzed simultaneously. In a preferred embodiment of the invention, system 70 includes tools for comparing gene tokens and/or their associated data and inference chains.

Another aspect of human interfacing arises when system 70 utilizes people as data sources. In a preferred embodiment of the invention, some data is preferentially received from a human source. Alternatively or additionally, some rules may require a human to decide on a particular matter. Alternatively or additionally, a human may be used as a secondary adapter for example, for operating various software tools, accessing data sources and/or aiding in parsing query results. Alternatively or additionally, in some cases system 70 may be stumped and require a human assistance or input, for example in developing a new scheme, in defining a data retrieval goal, in selecting appropriate databases and/or in other activities of the system.

In a preferred embodiment of the invention, system 70 performs a follow up on requests send to a user, providing the user with timely reminders.

In a preferred embodiment of the invention, system 70 may require data which does not exist in any data source. In a preferred embodiment of the invention, system 70 may formulate and/or send a work order to a laboratory work group.76. Alternatively or additionally, system 70 may employ an additional intelligent search agent, for example for searching the Internet.

In a preferred embodiment of the invention, when providing work to a human, system 70 includes an explanation on why the work is important, who the results might aid, an estimated difficulty and/or a priority level. It is expected that the inclusion of such items of information will assist in getting the work done.

Alternatively or additionally, to system 70 communicating with persons as tools, in a preferred embodiment of the invention, system 70 can directly or indirectly (via execution of other software programs) operate automated laboratory equipment. Automatic equipment includes, for example, robotic arms, flow-through chips and drug screening assay.

CONFIGURATION CONTROL

In a preferred embodiment of the invention, system 70 performs activities which do not directly relate to knowledge discovery. As shown in Fig. 4, system 70 may be connected to

many parts of an organization. In a preferred embodiment of the invention, some of the activities of system 70 are directed towards optimizing the contact and/or providing feedback to the parts of the organization. In one example, system 70 tracks (for example by using "reader" response messages) the effectiveness and/or suitability of communications to users. Such feedback, for example can be used to adjust the 'annoyance level' in a user's user model, as described above.

In another example, system 70 may perform QA/QC (quality assurance/quality control) functions. Typically, a significant amount of data generated by the organization may be used to feed system 70. Any problems with the data are preferably detected by system 70. Alternatively or additionally, system 70 includes inference rules for explaining inconsistency by indicating a problem with data. Also, system 70 can cross-correlate internal results with data from external databases. Also, system 70 can compare old conclusions against new data. The notification of appropriate personnel of these problems by system 70 has already been described above.

In a preferred embodiment of the invention, an activity of system 70 is resource allocation. In some cases, the only "noticeable" resource used by agents 102 is computer time. In addition however, system 70 may have other resources, for example money, for paying for data from fee-based data bases, work-hours and laboratory-hours, for performing work by people and/or machinery, communication bandwidth, disk storage space and time, for example deadlines or time windows where certain activities may be performed. Another type of resource is a database which can only be accessed one-at-a time.

In a preferred embodiment of the invention, resource allocation is reasoned over via the inference engine, and these resources are allocated based on priority between the gene tokens. Additionally or alternatively, each gene token may have associated therewith a upper spending limit and/or an upper spending limit per cycle. In a preferred embodiment of the invention, tokens are evaluated using a round-robin mechanism. Preferably however, a timer is set so that the data retrieval and knowledge discovery cycle is kept below a maximum duration. Alternatively or additionally, other resource allocation mechanisms may be used. For example, if work on a particular agent is hanging due to input being awaited, control may pass to a different token or a different agent on the same computer.

In a preferred embodiment of the invention, configuration control extends to data, information knowledge and/or rules. In a preferred embodiment of the invention, each gene token has stored therewith (perhaps in the gene token database) a snapshot of the current data mining / knowledge discovery state, allowing the agent to return to this state at a later time.

Alternatively or additionally, the gene token has stored therewith a link to and/or a copy of data and/or rules used to generate information in the gene token. In a preferred embodiment of the invention, if data and/or rules are determined to be erroneous and are corrected, deleted and/or otherwise updated, the gene tokens may be rolled back and re-evaluated. In a preferred embodiment of the invention, tracking includes also changes to data, rules and/or adapters, which changes are initiated by users. Such tracking is useful if a change in the configuration of system 70 had an unexpected effect on knowledge discovery, which effect is appreciated only at a later time.

In a preferred embodiment of the invention, one or more of the following mechanisms is used to detect that data has been updated:

(a) maintaining a mirror of the data and periodically comparing it to the database, or when a new version of the database is released;

(b) receiving a list of changes when a new version of a database becomes available;

(c) periodically checking data elements, preferably using a "staleness" indicator associated with data type, field type , gene identification and/or database identification; and/or

(d) randomly checking data elements.

In a preferred embodiment of the invention, the above mechanisms are implemented as inference rules.

In a preferred embodiment of the invention, system 70 includes some data security abilities, preferably in the form of granting users selective views of the data depending on their security level. Other security abilities preferably include limiting the ability of certain users to generate tasks, modify data and/or rules, affect the operation of the system for other users and/or access or expend certain resources. Thus, some users may not be able to view all the data. Alternatively or additionally, some data may be available, however certain users will not be notified of its availability. Alternatively or additionally, the degree of detail in messages may depend on the subject matter, recipient and/or transmission method. In some embodiments, data is completely compartmentalized. Alternatively, some data may be available across gene tokens for knowledge discovery purposes, even though not all the users will be able to view the data when analyzing a gene token.

In a preferred embodiment of the invention, system 70 includes an ability to self-monitor various aspects of its functioning. Preferably, one or more of the following aspects may be monitored:

(a) frequency of evaluation of rules;

(b) relative access to different databases;

(c) error rate in different databases;

(d) acceptance rate of inferences by users;

(e) data sources which do not respond (or a temporal profile of response times); and/or

(f) positive and/or negative contributions of data sources and/or of individuals to the success of knowledge discovery.

Preferably, the results of the self-monitoring are presented to an operator. Alternatively or additionally, some of these results may be used by the system to optimize its performance, for example by using erroneous databases less often and/or by using databases which had been previously neglected.

DEALING WITH ERRORS

In a preferred embodiment of the invention, any action or inferences taken by system 70 is logged, so that the source of errors (and good results) may be traced. Not only might this allow the system to analyze its past performance and induce new rules based on this analysis, but also preferably, when a user checks a gene token and/or a current state of a system, the user also checks the data used to reach the inferences – this is possibly in addition to checking the inference rules themselves. Often, a piece of data which is accepted by system 70 may be immediately identifiable as erroneous by a user. Alternatively or additionally, the user analyses the confidence level of the system in data, information and/or knowledge. Alternatively or additionally, a user can selectively view other portions of system 70, for example those which formulate queries, set data requirement goals and/or parse. If these portions are implemented as inference rules, a user can preferably selectively view only those inference rules. If these portions are not implemented as inference rules a user can preferably view source code, parameter settings and/or configuration files for these portions.

In a preferred embodiment of the invention, system 70 includes automatic error correction mechanisms. Preferably, when an error or an inconsistency is detected, system 70 attempts to determine where the error occurs, for example in which data source. Alternatively or additionally, system 70 isolates the error and freezes inferences involving the error, for example, at least until a user intervenes. In one example, if three data sources return sequence information, which do not match, a vote may be cast to select the correct sequence data. As a result, the contig and/or full-length sequence is preferably identified: Further inferences may yield a protein, protein family, protein homology and/or protein function.

In some cases, correctness may be determined by allowing the system to continue using the possibly erroneous data and then attempting to determine, at a later time, if indeed

the data was erroneous. In a preferred embodiment of the invention, agent 102 is split into a plurality of threads, each of which assumes that a different data element is correct. At a later time, the state of the different threads are compared to determine what, if any, difference may be found between them. Such thread splitting may also be used to perform probabilistic inferences, for example to evaluate a rule which provides several possible outcomes, possibly each with an associated probability. If the total probability is high enough, a separate thread may be formed, to test that possibility.

In a preferred embodiment of the invention, a message is sent to the data source indicating the error and, preferably, explaining how the error was caught.

One special type of error is when a data element does not match a conclusion. There can be two reasons, either one of the data element and the conclusion are erroneous or the data elements is being miss-construed. In a preferred embodiment of the invention, system 70 attempts to assert the second possibility by attempting a different interpretation of the inconsistent data. In one example, a cancer causing gene may be detected, but the gene token may have two different sequences. One possible explanation is that there are two types of cancer, each caused by possibly similar mechanisms, but as a result of different genes.

In a preferred embodiment of the invention, detected errors are used for quality control purposes, especially if the data source is within the organization.

LEARNING

In a preferred embodiment of the invention, system 70 is a learning system which adapts itself to a changing world. One important source of information is the above described self monitoring ability. Preferably, the results of the learning are applied towards modifying the system configuration and/or its inference rules. In a preferred embodiment of the invention, one or more of the following may be learned by system 70:

(a) which databases are to be preferred;

(b) semantic terms used in a particular database;

(c) types and distribution of mistakes in certain databases;

(d) personal preferences; and/or

(e) preferred order of rule application.

In a preferred embodiment of the invention, learning is implemented by modifying system configuration tables and/or by setting various system parameters. Alternatively or additionally, learning may be implemented by erasing rules, modifying rule matching stringency and/or modifying rule application priority.

In a preferred embodiment of the invention, system 70 may perform testing activities, to compare two variants of a rule, parameter setting, data interpretation, threshold value and/or other modifiable items in system 70. Preferably, system 70 tests a hypothesis, regarding which variant is to be preferred, by executing two or more variants, preferably simultaneously. The variant which achieved the better results or the variant which did not cause any (or less) inconsistency in stored content is preferably considered to be the better variant.

SOME POSSIBLE MODIFICATIONS

Although many functions of system 70 may be achieved using inference rules, in some preferred embodiments of the invention, these functions may be achieved using other means. Such means include scripts, preferably in an interpreted language; production rules with associated databases of situations; and/or procedural software components. As described above, these functions may also include any of the steps of Fig. 1. In a preferred embodiment of the invention, these non-inference elements may be used to generate a model of biological activity, as the data is being accumulated and verified.

In a preferred embodiment of the invention, system 70 includes a current operating state, which may affect and/or be affected by the inference rules. For example, the number of discovered drug candidates may affect the strictness in which inference rules are applied and/or in which queries are broadened. Preferably, different data matching functions are provided as a function of the system operating state.

In a preferred embodiment of the invention, system 70 execute multiple inter-communicating agents simultaneously. Such a configuration may be useful if for example each agent attacks a different aspect of a same problem. In a preferred embodiment of the invention, the agents communicate when one agent discovers knowledge relevant to another agent. Preferably communication uses the KQML standard for communications, using either CLIPS or KIF as an encapsulated knowledge representation language.

In a preferred embodiment of the invention, multi-agent system 70 includes a facilitator agent (or adapter), for example as defined in the FIPA (Foundation for Intelligent Physical Agents) guidelines. This facilitator agent may be used to enable one agent to find a second agent have a specialty of finding a certain type of data. Thus, an agent may function as a data source.

In a preferred embodiment of the invention, system 70 includes a communication coordinator which coordinates the communications of the multiple agents with the external and/or internal databases.

The present invention has been described with respect to a genomic information system. However, it should be noted that the ideas and/or features of the system may also be applied to other fields. In particular other fields of biology suffer from the above described problems in genomics: scale, updating, errors, heterogeneity and complexity, albeit possibly not to the same degree of severity as genomics. One example is proteomics. Another example is genetic research. It should also be appreciated that the genomic applications may be expanded to incorporate other types of information associated with the drug development process, for example for lead compound discovery and toxicology testing. It is also possible to apply some of the above described embodiments of the invention to non-biological uses, for example for industrial intelligence and financial information. In these application, the nucleation center will not be a gene token but may be, for example, a "corporation token". Additionally, biological relevance and biological type schemes will be replaced by "financial relevance" and financial type schemes, which may embody the type of information that is expected to be exist for a corporation.. However, it is noted that the types of problems in those fields are somewhat different and possibly easier to solve than in genomics, especially with regard to scale, complexity, heterogeneity and error rate.

The present invention has been described in terms of preferred, non-limiting embodiments thereof. It should be understood that features described with respect to one embodiment may be used with other embodiments and that not all embodiments of the invention have all of the features shown in a particular figure. In particular, the scope of the invention is not defined by the preferred embodiments but by the following claims. Section titles, where they appear are not to be construed in limiting subject matter described therein, rather section titles are meant only as an aid in browsing this specification. When used in the following claims, the terms "comprises", "comprising", "includes", "including" or the like means "including but not limited to".

CLAIMS

1. A method of genomic data discovery, comprising:

(a) providing a gene data base comprising at least 10 genes;

(b) selecting one of said at least 10 genes;

(c) discovering knowledge for said selected gene;

(d) repeating said (b) and (c) for a plurality of said genes; and

(e) repeating said (b)-(d) a plurality of times such that knowledge is discovered substantially in parallel for all the selected genes.

2. A method according to claim 1, wherein said (b)-(e) are performed substantially without human intervention.

3. A method according to any of claims 1-2, comprising evaluating, by a computer and without requiring additional input from an operator, said genes for which knowledge has been discovered.

4. A method according to claim 3, wherein automatically evaluating said genes comprises ranking said genes according to their suitability for being drug leads.

5. A method according to any of claims 3-4, comprising deciding on further selecting of said genes in (b), responsive to said evaluation.

6. A method according to any of claims 1-5, wherein (c) comprises determining data needs for said genes.

7. A method according to claim 6, wherein each of said genes is associated with a scheme and wherein determining data needs comprise analyzing said scheme to determine data needs.

8. A method according to any of claims 6-7, wherein (c) comprises formulating queries to obtain said needed data.

9. A method according to any of claims 6-8, wherein (c) comprises setting up sub-goals for obtaining said data needs.

10. A method according to any of claims 6-9, comprising, selecting suitable databases for said data needs.

11. A method according to claim 10, wherein selecting suitable databases comprises selecting at least 10 databases for at least 5 needed elements of data.

12. A method according to any of claims 1-11, wherein said (c) comprises receiving data for a plurality of data sources.

13. A method according to claim 12, comprising integrating said received data.

14. A method according to claim 12 or claim 13, comprising analyzing said received data to produce knowledge.

15. A method according to claim 14, comprising returning said selected gene to said database, in association with said discovered knowledge.

16. A method according to any of claims 1-15, wherein said at least 10 genes comprises at least 50 genes.

17. A method according to any of claims 1-15, wherein said at least 10 genes comprises at least 100 genes.

18. A method according to any of claims 1-15, wherein said at least 10 genes comprises at least 300 genes.

19. A method according to any of claims 1-15, wherein said plurality of genes comprises at least 20% of said at least 10 genes.

20. A method according to any of claims 1-15, wherein said plurality of genes comprises at least 50% of said at least 10 genes.

21. A method according to any of claims 1-15, wherein said plurality of genes comprises at least 80% of said at least 10 genes.

22. A method of genomic knowledge discovery, comprising:
determining at least one required data element for at least one gene;
querying a plurality of at least 50 databases for said at least one required data element;
receiving responses from said databases; and
analyzing said responses to increase knowledge for said at least one gene.

23. A method according to claim 22, wherein said at least 50 databases comprise at least 100 databases.

24. A method according to claim 22, wherein said at least 50 databases comprise at least 300 databases.

25. A method according to any of claims 22-24, wherein said databases are all queried for a same data value.

26. A method according to any of claims 22-25, wherein said method is performed , by a computer and without requiring additional input from an operator,.

27. A method of automated knowledge discovery, comprising:
continuously operating a knowledge discovery cycle comprising:
querying a database to receive data; and
performing inferences on said data to generate knowledge; and
re-evaluating said inferences when said database is modified

28. A method according to claim 27, wherein said cycle is continuously operated over one week.

29. A method according to claim 27, wherein said cycle is continuously operated over one month.

30. A method according to claim 27, wherein said cycle is continuously operated over six months.

31. A method of genomic knowledge discovery, comprising:

(a) selecting a gene token;

(b) determining data requirements for said gene token;

(c) requesting and receiving data responsive to said data requirements;

(d) analyzing said received information to increase knowledge for said gene token; and

(e) repeating (b)-(d) for the same gene token, at least 50 times.

32. A method according to claim 31, wherein at least 50 times comprises at least 100 times.

33. A method according to claim 31, wherein at least 50 times comprises at least 200 times.

34. A knowledge discovers system comprising:

a first unit for determining data needs and analyzing data returned responsive to said needs; and

a plurality of at least 10 adapter units, for accessing at least 10 dissimilar data sources to provide said needed data.

35. A system according to claim 34, wherein said first unit comprises an inference engine comprising rules for analyzing biological data.

36. A system according to claim 35, wherein said biological data comprises genomic data.

37. A system according to any of claims 34-36, wherein said at least 10 adapter units comprise at least 50 adapter units for 50 dissimilar data sources.

38. A system according to any of claims 34-36, wherein said at least 10 adapter units comprise at least 100 adapter units for 100 dissimilar data sources.

39. A system according to any of claims 34-36, wherein said at least 10 adapter units comprise at least 300 adapter units for 300 dissimilar data sources.

40. A system according to any of claims 34-39, wherein said at least 10 data sources comprise at least 10 data analysis tools.

41. A system according to any of claims 34-39, wherein said at least 10 data sources comprise at least 30 data analysis tools.

42. A system according to any of claims 34-41, wherein said adapters are programmed in an interpreted text processing language.

43. A system according to claim 42, wherein said adapters are programmed in language including classes.

44. A system according to claim 42 or claim 43, wherein said adapters are programmed in a Perl-like language.

45. A system according to any of claims 34-43, comprising a central adapter registry, wherein, each of said adapter registers its availability in said central registry.

46. A system according to claim 45, wherein said registry is implemented as assertions.

47. A system according to claim 45 or claim 46, wherein said adapters register their data provision capabilities in said registry.

48. A system according to any of claims 34-47, wherein said first unit analyses said data requirements to determine selected ones of said data sources to query.

49. A method of ranking genes for an application, comprising:

providing a plurality of gene tokens;

applying, by a computer and without requiring additional input from an operator, application-specific ranking rules to said gene tokens; and

ranking the gene tokens responsive to said application of rules.

50. A method according to claim 49, wherein said plurality of gene tokens comprise at least 10 gene tokens.

51. A method according to claim 49, wherein said plurality of gene tokens comprise at least 30 gene tokens.

52. A method according to claim 49, wherein said plurality of gene tokens comprise at least 50 gene tokens.

53. A method according to any of claims 49-52, wherein said application comprises drug discovery.

54. A method according to claim 53, wherein said application specific rules comprise biological relevance rules.

55. A method according to claim 54, wherein said biological relevance rules comprise rules which match said gene tokens to a disease model across a plurality of biological dimensions.

56. A method according to any of claims 53-55, wherein said application specific rules comprise rules which determine experimental tractability.

57. A method according to any of claims 53-55, wherein said application specific rules comprise rules which determine pharmaceutical tractability.

58. A method according to claim 57, wherein pharmaceutical tractability determination comprises an indication of ease of finding an effective pharmaceutical.

59. A method according to claim 57, wherein pharmaceutical tractability determination comprises an indication of ease of finding a pharmaceutical with a low level of side effects.

60. A method of genomic information analysis, comprising:

providing a first model of a biological relationship which interrelates a first plurality of genes or proteins;

providing a second model of a biological relationship which interrelates a second plurality of genes or proteins; and

applying inference rules to said first and second models to infer missing information.

61. A method according to claim 60, wherein said applying comprises determining data needs for at least one of said models, based on said applied inference rules.

62. A method according to claim 60 or claim 61, wherein said biological relationships comprise enzymatic pathways.

63. A method according to any of claims 60-62, wherein said biological relationships are in different species.

64. A method of automated genomic knowledge discovery, comprising:

analyzing a gene token to determine required data;

selecting at least one suitable human expert; and

querying the at least one selected human expert for the required data.

65. A method of automated genomic knowledge discovery, comprising:

analyzing a gene token to determine required data; and

generating, by a computer and without requiring additional input from an operator, a work order to a laboratory to generate the required data.

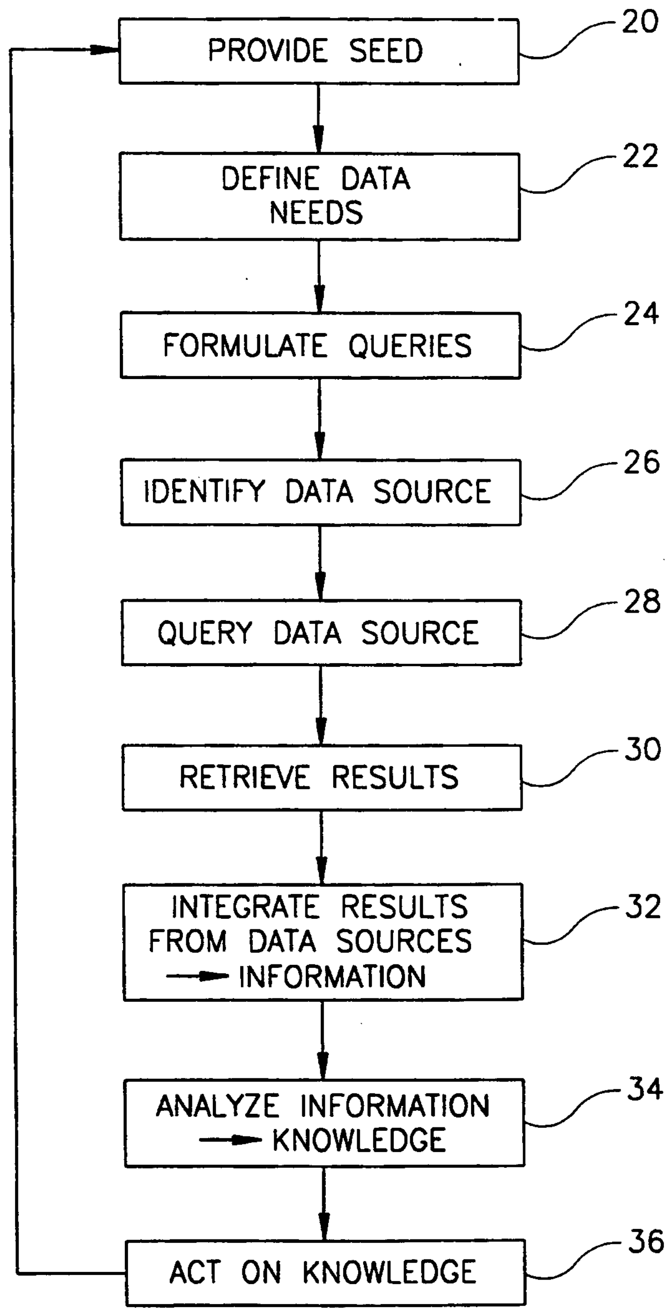

PROVIDE SEED
20
DEFINE DATA NEEDS
22
FORMULATE QUERIES
24
IDENTIFY DATA SOURCE
26
QUERY DATA SOURCE
28
RETRIEVE RESULTS
30
INTEGRATE RESULTS FROM DATA SOURCES → INFORMATION
32
ANALYZE INFORMATION → KNOWLEDGE
34
ACT ON KNOWLEDGE
36

FIG. 1

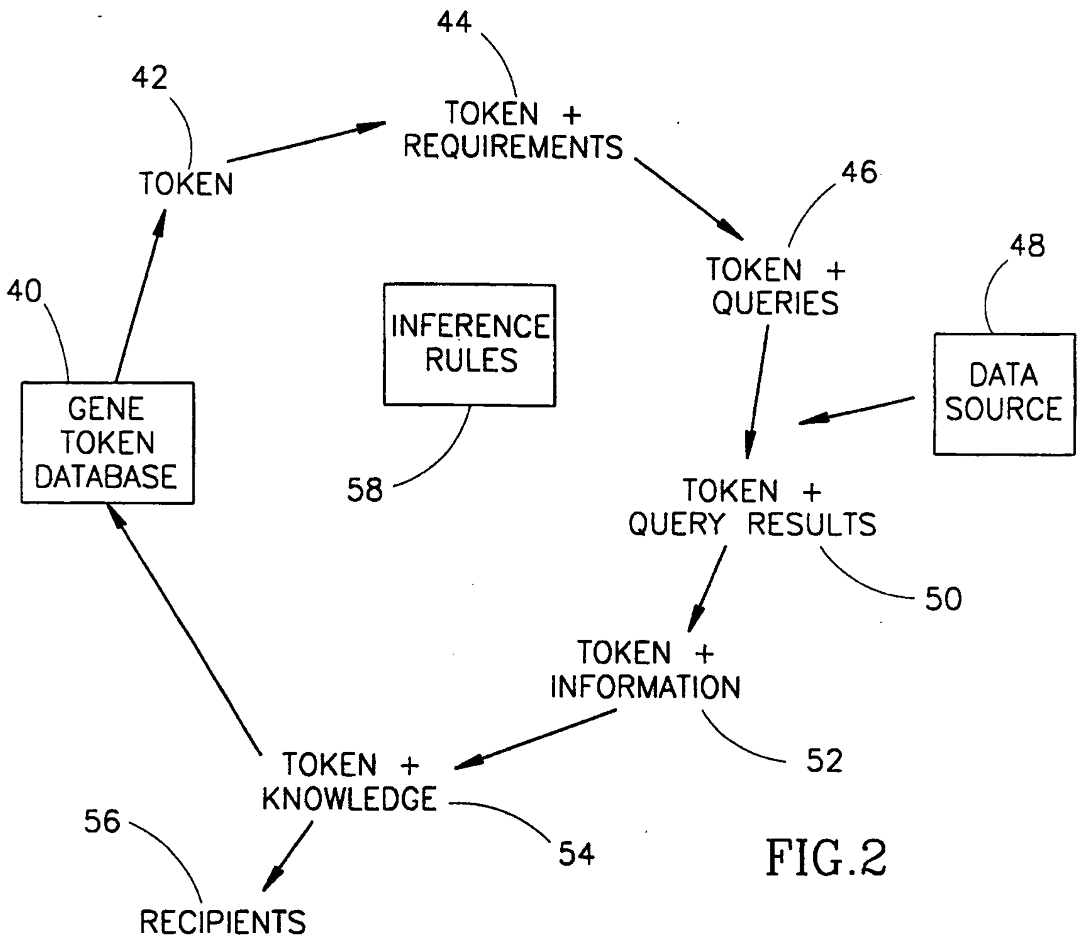

44
42
TOKEN + REQUIREMENTS
TOKEN
46
48
40
TOKEN + QUERIES
INFERENCE RULES
DATA SOURCE
GENE TOKEN DATABASE
58
TOKEN + QUERY RESULTS
50
TOKEN + INFORMATION
52
TOKEN + KNOWLEDGE
56
54
RECIPIENTS

FIG.2

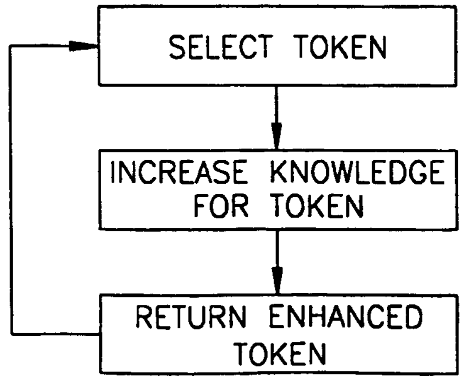

SELECT TOKEN
INCREASE KNOWLEDGE FOR TOKEN
RETURN ENHANCED TOKEN

FIG.3

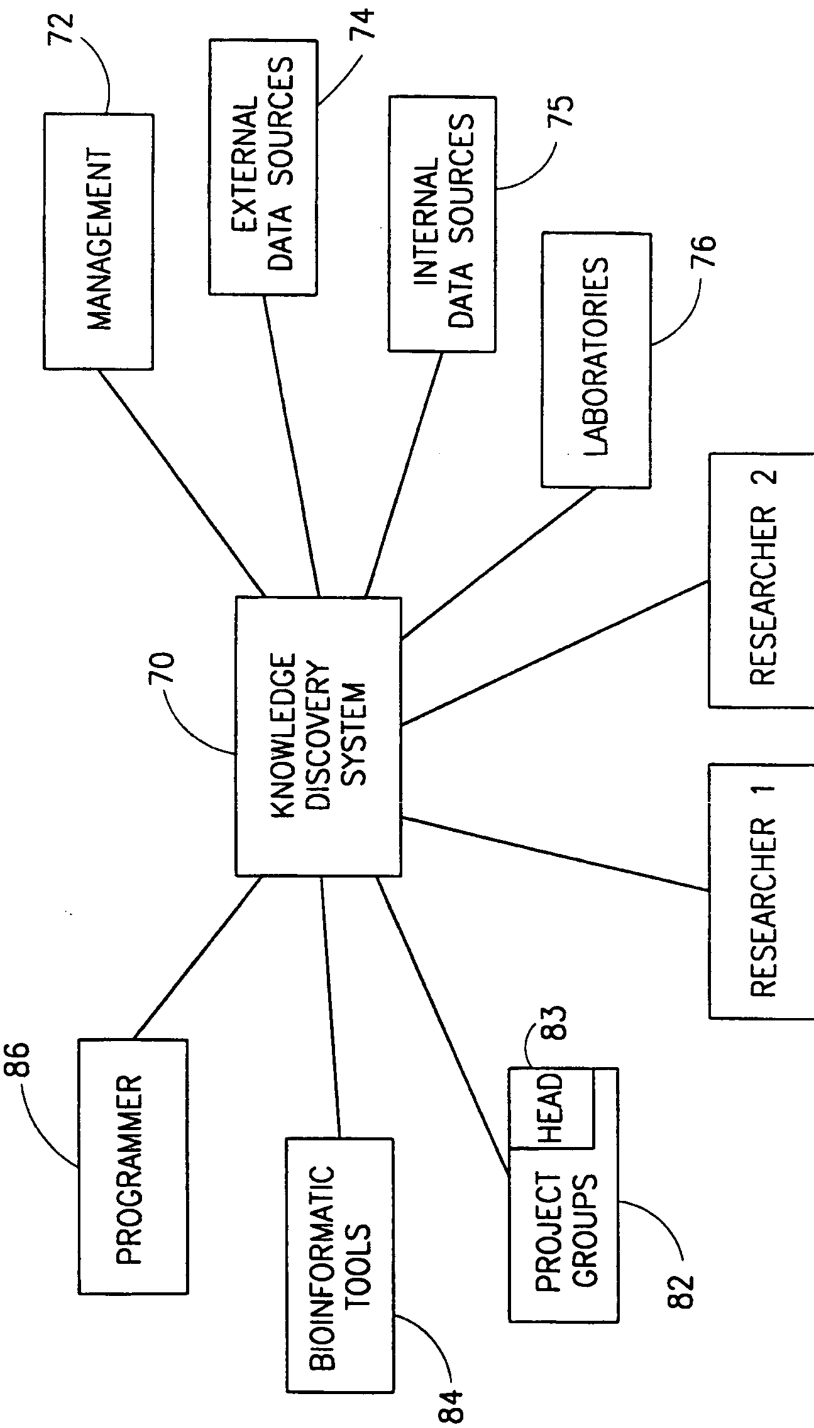

MANAGEMENT
72
EXTERNAL DATA SOURCES
74
INTERNAL DATA SOURCES
75
LABORATORIES
76
RESEARCHER 2
80
KNOWLEDGE DISCOVERY SYSTEM
70
RESEARCHER 1
78
PROGRAMMER
86
BIOINFORMATIC TOOLS
84
PROJECT GROUPS
82
HEAD
83

FIG.4

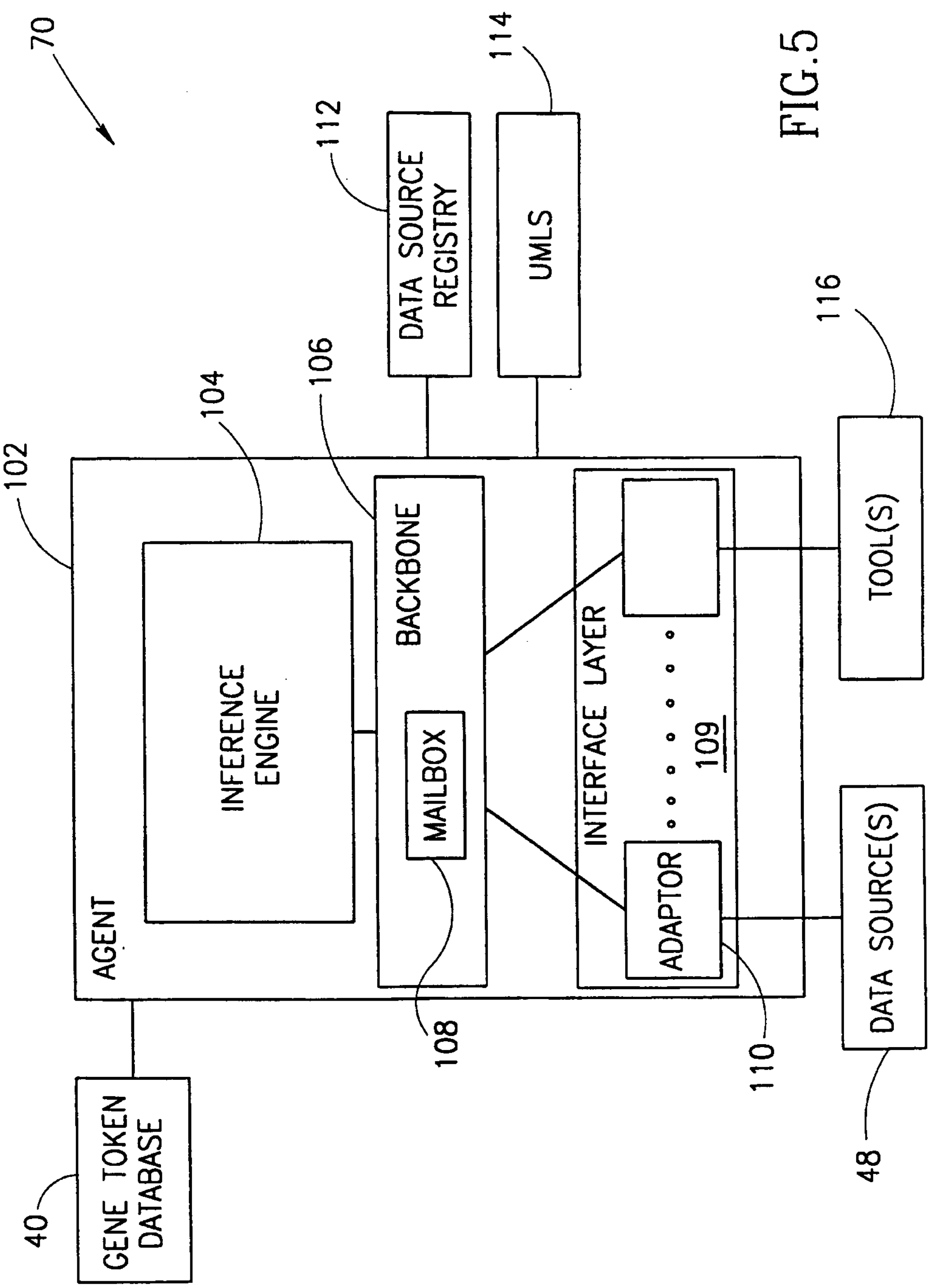

70
102
AGENT
104
INFERENCE ENGINE
106
BACKBONE
MAILBOX
108
INTERFACE LAYER
109
ADAPTOR
110
112
DATA SOURCE REGISTRY
114
UMLS
116
TOOL(S)
DATA SOURCE(S)
48
40
GENE TOKEN DATABASE

FIG.5